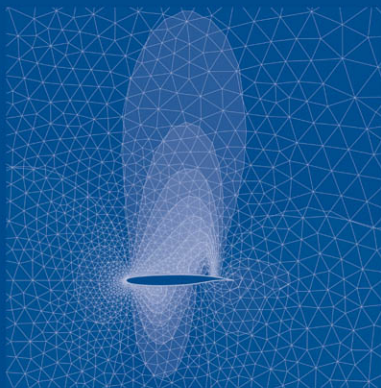


NUMERICAL MATHEMATICS  
AND SCIENTIFIC COMPUTATION

# Numerical Analysis and Optimization

An Introduction to Mathematical  
Modelling and Numerical Simulation

GRÉGOIRE ALLAIRE



OXFORD SCIENCE PUBLICATIONS

NUMERICAL MATHEMATICS AND SCIENTIFIC COMPUTATION

---

*Series Editors*

G.H. GOLUB, A. GREENBAUM  
A.M. STUART, E. SÜLI

## NUMERICAL MATHEMATICS AND SCIENTIFIC COMPUTATION

---

### *Books in the series*

Monographs marked with an asterisk (\*) appeared in the series ‘Monographs in Numerical Analysis’ which has been folded into, and is continued by, the current series.

\* P. Dierckx: *Curve and surface fittings with splines*

\* J.H. Wilkinson: *The algebraic eigenvalue problem*

\* I. Duff, A. Erisman, and J. Reid: *Direct methods for sparse matrices*

\* M.J. Baines: *Moving finite elements*

\* J.D. Pryce: *Numerical solution of Sturm–Liouville problems*

K. Burrage: *Parallel and sequential methods for ordinary differential equations*

Y. Censor and S.A. Zenios: *Parallel optimization: theory, algorithms and applications*

M. Ainsworth, J. Levesley, W. Light, and M. Marletta: *Wavelets, multilevel methods and elliptic PDEs*

W. Freeden, T. Gervens, and M. Schreiner: *Constructive approximation on the sphere: theory and applications to geomathematics*

C.H. Schwab: *p- and hp- finite element methods: theory and applications to solid and fluid mechanics*

J.W. Jerome: *Modelling and computation for applications in mathematics, science, and engineering*

Alfio Quarteroni and Alberto Valli: *Domain decomposition methods for partial differential equations*

G.E. Karniadakis and S.J. Sherwin: *Spectral/hp element methods for CFD*

I. Babuška and T. Strouboulis: *The finite element method and its reliability*

B. Mohammadi and O. Pironneau: *Applied shape optimization for fluids*

S. Succi: *The Lattice Boltzmann Equation for fluid dynamics and beyond*

P. Monk: *Finite element methods for Maxwell’s equations*

A. Bellen and M. Zennaro: *Numerical methods for delay differential equations*

J. Modersitzki: *Numerical methods for image registration*

M. Feistauer, J. Felcman, and I. Straškraba: *Mathematical and computational methods for compressible flow*

W. Gautschi: *Orthogonal polynomials: computation and approximation*

M.K. Ng: *Iterative methods for Toeplitz systems*

Michael Metcalf, John Reid, and Malcolm Cohen: *Fortran 95/2003 explained*

George Em Karniadakis and Spencer Sherwin: *Spectral/hp element methods for CFD, second edition*

Dario A. Bini, Guy Latouche, and Beatrice Meini: *Numerical methods for structured Markov chains*

Howard Elman, David Silvester, and Andy Wathen: *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*

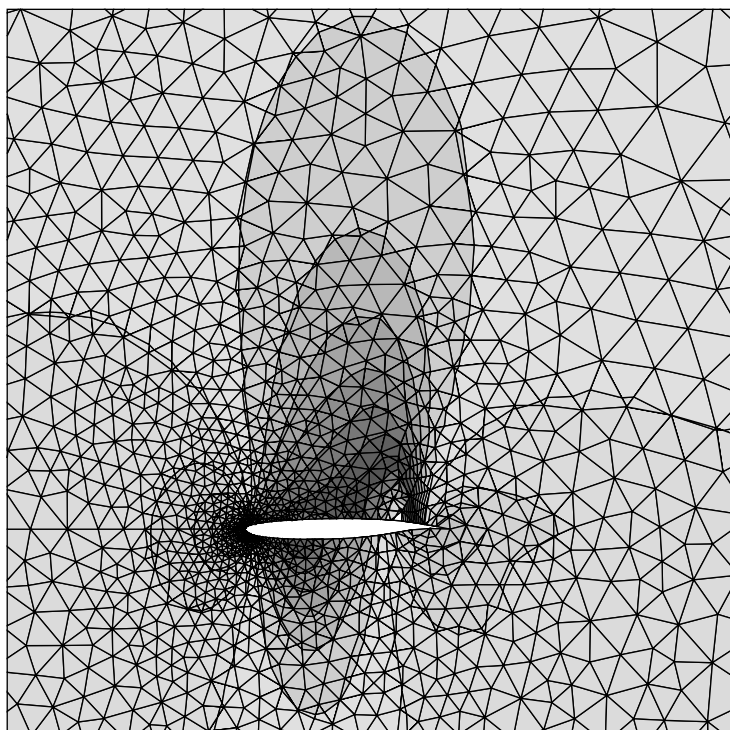
Moody Chu and Gene Golub: *Inverse eigenvalue problems: theory and applications*

Jean-Frédéric Gerbeau, Claude Le Bris, and Tony Lelièvre: *Mathematical methods for the magnetohydrodynamics of liquid metals*

Grégoire Allaire: *Numerical analysis and optimization*

# NUMERICAL ANALYSIS AND OPTIMIZATION

An introduction to mathematical modelling  
and numerical simulation



Grégoire Allaire  
*École Polytechnique*  
Translated by Dr Alan Craig  
University of Durham

**OXFORD**  
UNIVERSITY PRESS

OXFORD

UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.  
It furthers the University's objective of excellence in research, scholarship,  
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi  
Kuala Lumpur Madrid Melbourne Mexico City Nairobi  
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece  
Guatemala Hungary Italy Japan Poland Portugal Singapore  
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press  
in the UK and in certain other countries

Published in the United States  
by Oxford University Press Inc., New York

© Grégoire Allaire 2007

The moral rights of the author have been asserted  
Database right Oxford University Press (maker)

First published 2007

All rights reserved. No part of this publication may be reproduced,  
stored in a retrieval system, or transmitted, in any form or by any means,  
without the prior permission in writing of Oxford University Press,  
or as expressly permitted by law, or under terms agreed with the appropriate  
reprographics rights organization. Enquiries concerning reproduction  
outside the scope of the above should be sent to the Rights Department,  
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover  
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloging in Publication Data

Data available

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India

Printed in Great Britain

on acid-free paper by

Biddles Ltd., King's Lynn, Norfolk

ISBN 978-0-19-920521-9 (Hbk.)

ISBN 978-0-19-920522-6 (Pbk.)

1 3 5 7 9 10 8 6 4 2

# Contents

<b>1</b>	<b>Introduction</b>	<b>ix</b>
<b>1</b>	<b>Introduction to mathematical modelling and numerical simulation</b>	<b>1</b>
1.1	General introduction . . . . .	1
1.2	An example of modelling . . . . .	2
1.3	Some classical models . . . . .	9
1.3.1	The heat flow equation . . . . .	9
1.3.2	The wave equation . . . . .	9
1.3.3	The Laplacian . . . . .	11
1.3.4	Schrödinger's equation . . . . .	12
1.3.5	The Lamé system . . . . .	12
1.3.6	The Stokes system . . . . .	13
1.3.7	The plate equations . . . . .	13
1.4	Numerical calculation by finite differences . . . . .	14
1.4.1	Principles of the method . . . . .	14
1.4.2	Numerical results for the heat flow equation . . . . .	17
1.4.3	Numerical results for the advection equation . . . . .	21
1.5	Remarks on mathematical models . . . . .	25
1.5.1	The idea of a well-posed problem . . . . .	25
1.5.2	Classification of PDEs . . . . .	28
<b>2</b>	<b>Finite difference method</b>	<b>31</b>
2.1	Introduction . . . . .	31
2.2	Finite differences for the heat equation . . . . .	32
2.2.1	Various examples of schemes . . . . .	32
2.2.2	Consistency and accuracy . . . . .	35
2.2.3	Stability and Fourier analysis . . . . .	36
2.2.4	Convergence of the schemes . . . . .	42
2.2.5	Multilevel schemes . . . . .	44
2.2.6	The multidimensional case . . . . .	46
2.3	Other models . . . . .	51
2.3.1	Advection equation . . . . .	51
2.3.2	Wave equation . . . . .	59

<b>3</b>	<b>Variational formulation of elliptic problems</b>	<b>65</b>
3.1	Generalities . . . . .	65
3.1.1	Introduction . . . . .	65
3.1.2	Classical formulation . . . . .	66
3.1.3	The case of a space of one dimension . . . . .	67
3.2	Variational approach . . . . .	68
3.2.1	Green's formulas . . . . .	68
3.2.2	Variational formulation . . . . .	71
3.3	Lax–Milgram theory . . . . .	73
3.3.1	Abstract framework . . . . .	73
3.3.2	Application to the Laplacian . . . . .	76
<b>4</b>	<b>Sobolev spaces</b>	<b>79</b>
4.1	Introduction and warning . . . . .	79
4.2	Square integrable functions and weak differentiation . . . . .	80
4.2.1	Some results from integration . . . . .	80
4.2.2	Weak differentiation . . . . .	81
4.3	Definition and principal properties . . . . .	84
4.3.1	The space $H^1(\Omega)$ . . . . .	84
4.3.2	The space $H_0^1(\Omega)$ . . . . .	88
4.3.3	Traces and Green's formulas . . . . .	89
4.3.4	A compactness result . . . . .	94
4.3.5	The spaces $H^m(\Omega)$ . . . . .	96
4.4	Some useful extra results . . . . .	98
4.4.1	Proof of the density theorem 4.3.5 . . . . .	98
4.4.2	The space $H(\operatorname{div})$ . . . . .	101
4.4.3	The spaces $W^{m,p}(\Omega)$ . . . . .	102
4.4.4	Duality . . . . .	103
4.5	Link with distributions . . . . .	105
<b>5</b>	<b>Mathematical study of elliptic problems</b>	<b>109</b>
5.1	Introduction . . . . .	109
5.2	Study of the Laplacian . . . . .	109
5.2.1	Dirichlet boundary conditions . . . . .	109
5.2.2	Neumann boundary conditions . . . . .	116
5.2.3	Variable coefficients . . . . .	123
5.2.4	Qualitative properties . . . . .	126
5.3	Solution of other models . . . . .	136
5.3.1	System of linear elasticity . . . . .	136
5.3.2	Stokes equations . . . . .	144
<b>6</b>	<b>Finite element method</b>	<b>149</b>
6.1	Variational approximation . . . . .	149
6.1.1	Introduction . . . . .	149
6.1.2	General internal approximation . . . . .	150
6.1.3	Galerkin method . . . . .	153
6.1.4	Finite element method (general principles) . . . . .	153
6.2	Finite elements in $N = 1$ dimension . . . . .	154

6.2.1	$\mathbb{P}_1$ finite elements . . . . .	154
6.2.2	Convergence and error estimation . . . . .	159
6.2.3	$\mathbb{P}_2$ finite elements . . . . .	163
6.2.4	Qualitative properties . . . . .	165
6.2.5	Hermite finite elements . . . . .	168
6.3	Finite elements in $N \geq 2$ dimensions . . . . .	171
6.3.1	Triangular finite elements . . . . .	171
6.3.2	Convergence and error estimation . . . . .	184
6.3.3	Rectangular finite elements . . . . .	191
6.3.4	Finite elements for the Stokes problem . . . . .	195
6.3.5	Visualization of the numerical results . . . . .	201
<b>7</b>	<b>Eigenvalue problems</b>	<b>205</b>
7.1	Motivation and examples . . . . .	205
7.1.1	Introduction . . . . .	205
7.1.2	Solution of nonstationary problems . . . . .	206
7.2	Spectral theory . . . . .	208
7.2.1	Generalities . . . . .	209
7.2.2	Spectral decomposition of a compact operator . . . . .	210
7.3	Eigenvalues of an elliptic problem . . . . .	213
7.3.1	Variational problem . . . . .	213
7.3.2	Eigenvalues of the Laplacian . . . . .	218
7.3.3	Other models . . . . .	221
7.4	Numerical methods . . . . .	224
7.4.1	Discretization by finite elements . . . . .	224
7.4.2	Convergence and error estimates . . . . .	227
<b>8</b>	<b>Evolution problems</b>	<b>231</b>
8.1	Motivation and examples . . . . .	231
8.1.1	Introduction . . . . .	231
8.1.2	Modelling and examples of parabolic equations . . . . .	232
8.1.3	Modelling and examples of hyperbolic equations . . . . .	233
8.2	Existence and uniqueness in the parabolic case . . . . .	234
8.2.1	Variational formulation . . . . .	234
8.2.2	A general result . . . . .	236
8.2.3	Applications . . . . .	241
8.3	Existence and uniqueness in the hyperbolic case . . . . .	246
8.3.1	Variational formulation . . . . .	246
8.3.2	A general result . . . . .	247
8.3.3	Applications . . . . .	249
8.4	Qualitative properties in the parabolic case . . . . .	253
8.4.1	Asymptotic behaviour . . . . .	253
8.4.2	The maximum principle . . . . .	255
8.4.3	Propagation at infinite velocity . . . . .	256
8.4.4	Regularity and regularizing effect . . . . .	257
8.4.5	Heat equation in the entire space . . . . .	259
8.5	Qualitative properties in the hyperbolic case . . . . .	261
8.5.1	Reversibility in time . . . . .	261



8.5.2	Asymptotic behaviour and equipartition of energy . . . . .	262
8.5.3	Finite velocity of propagation . . . . .	263
8.6	Numerical methods in the parabolic case . . . . .	264
8.6.1	Semidiscretization in space . . . . .	264
8.6.2	Total discretization in space-time . . . . .	266
8.7	Numerical methods in the hyperbolic case . . . . .	269
8.7.1	Semidiscretization in space . . . . .	270
8.7.2	Total discretization in space-time . . . . .	271
<b>9</b>	<b>Introduction to optimization</b>	<b>277</b>
9.1	Motivation and examples . . . . .	277
9.1.1	Introduction . . . . .	277
9.1.2	Examples . . . . .	278
9.1.3	Definitions and notation . . . . .	284
9.1.4	Optimization in finite dimensions . . . . .	285
9.2	Existence of a minimum in infinite dimensions . . . . .	287
9.2.1	Examples of nonexistence . . . . .	287
9.2.2	Convex analysis . . . . .	289
9.2.3	Existence results . . . . .	292
<b>10</b>	<b>Optimality conditions and algorithms</b>	<b>297</b>
10.1	Generalities . . . . .	297
10.1.1	Introduction . . . . .	297
10.1.2	Differentiability . . . . .	298
10.2	Optimality conditions . . . . .	303
10.2.1	Euler inequalities and convex constraints . . . . .	303
10.2.2	Lagrange multipliers . . . . .	306
10.3	Saddle point, Kuhn–Tucker theorem, duality . . . . .	317
10.3.1	Saddle point . . . . .	317
10.3.2	The Kuhn–Tucker theorem . . . . .	318
10.3.3	Duality . . . . .	320
10.4	Applications . . . . .	323
10.4.1	Dual or complementary energy . . . . .	323
10.4.2	Optimal command . . . . .	326
10.4.3	Optimization of distributed systems . . . . .	330
10.5	Numerical algorithms . . . . .	332
10.5.1	Introduction . . . . .	332
10.5.2	Gradient algorithms (case without constraints) . . . . .	333
10.5.3	Gradient algorithms (case with constraints) . . . . .	336
10.5.4	Newton’s method . . . . .	342
<b>11</b>	<b>Methods of operational research</b>	
	(Written in collaboration with Stéphane Gaubert)	<b>347</b>
11.1	Introduction . . . . .	347
11.2	Linear programming . . . . .	348
11.2.1	Definitions and properties . . . . .	348
11.2.2	The simplex algorithm . . . . .	353
11.2.3	Interior point algorithms . . . . .	357

11.2.4 Duality . . . . .	358
11.3 Integer polyhedra . . . . .	361
11.3.1 Extreme points of compact convex sets . . . . .	362
11.3.2 Totally unimodular matrices . . . . .	364
11.3.3 Flow problems . . . . .	368
11.4 Dynamic programming . . . . .	371
11.4.1 Bellman's optimality principle . . . . .	372
11.4.2 Finite horizon problem . . . . .	372
11.4.3 Minimum cost path, or optimal stopping, problem . . . . .	375
11.5 Greedy algorithms . . . . .	380
11.5.1 General points about greedy methods . . . . .	380
11.5.2 Kruskal's algorithm for the minimum spanning tree problem . . . . .	381
11.6 Separation and relaxation . . . . .	383
11.6.1 Separation and evaluation (branch and bound) . . . . .	383
11.6.2 Relaxation of combinatorial problems . . . . .	388
<b>12 Appendix Review of hilbert spaces</b>	<b>399</b>
<b>13 Appendix Matrix Numerical Analysis</b>	<b>405</b>
13.1 Solution of linear systems . . . . .	405
13.1.1 Review of matrix norms . . . . .	406
13.1.2 Conditioning and stability . . . . .	409
13.1.3 Direct methods . . . . .	411
13.1.4 Iterative methods . . . . .	424
13.1.5 The conjugate gradient method . . . . .	428
13.2 Calculation of eigenvalues and eigenvectors . . . . .	435
13.2.1 The power method . . . . .	436
13.2.2 The Givens–Householder method . . . . .	438
13.2.3 The Lanczos method . . . . .	442
<b>Index</b>	<b>451</b>
<b>Index notations</b>	<b>455</b>

*This page intentionally left blank*

*To the memory of Jacques-Louis LIONS (1928–2001)*  
*Professor at the École Polytechnique from 1966 to 1986*

*This page intentionally left blank*

# Introduction

This course treats two essential subjects, among many others, in applied mathematics: numerical analysis and optimization. Before even presenting these two disciplines, let us immediately say that through their teaching the objective of this course is to introduce the reader to the world of **mathematical modelling** and **numerical simulation** which have gained considerable importance in these last decades in all areas of science and industrial applications (or engineering science). Mathematical modelling is the art (or the science, depending on the point of view) of representing (or transforming) a physical reality into abstract models which are accessible to analysis and to calculation. Numerical simulation is, of course, the process which allows us to calculate the solutions of these models on a computer, and thus to simulate physical reality.

But, first of all, what is applied mathematics? To say that it is mathematics turned towards applications would be a tautology and a false characterization. In effect, throughout time, mathematicians have been inspired by the practical problems that they have tried to solve, however, the emergence of applied mathematics as an independent discipline is relatively recent. In fact, everything changed with the appearance of the first computers shortly after the Second World War. More than for any other discipline the computer has been a revolution for mathematics: in effect it has opened up a new field, that of modelling and simulation. The computer has made mathematics an experimental science (we make ‘numerical experiments’ as others make physical experiments), and the design, as well as the analysis of methods of calculation on a computer, has become a new branch of mathematics: this is numerical simulation. This progress also made it possible for mathematics to attack much more complex and concrete problems, resulting from immediate industrial or scientific motivations, to which we can bring both qualitative and quantitative responses: this is mathematical modelling.

We can thus characterize applied mathematics as the mathematics of modelling and numerical simulation. From this point of view, applied mathematics lies at the intersection of many scientific disciplines: mathematics, computing, physical sciences, chemistry, mechanics, biology, economics, and engineering sciences (under this last term we usually group the different fields of industrial applications such as aeronautics, power generation, finance, etc.). The American mathematician Joseph Keller said as a joke that ‘pure mathematics is a branch of applied mathematics’. He wanted to highlight the multidisciplinary character of applied mathematics (but it is not excluded that he also wanted to pay back some ‘pure’ mathematicians who affect to despise applied mathematics).

Paraphrasing the title of a famous film, my colleague Pierre-Louis Lions claims that applied mathematics is characterized by three things: *Sex, Lies, and Videotapes*. Videocassettes are of course the symbols of digital simulation (and of the films that they produce), lies

correspond to models (not always faithful to reality), and the sex is obviously mathematical analysis (inextinguishable engine of human passions and source of much pleasure).

After this (long) detour we can now return to the title of this course. Numerical analysis is thus the discipline which conceives and analyses the methods or algorithms of numerical calculation. In addition optimization is the theory of methods which allow us to improve the operation, output, or the response of a system by maximizing or minimizing associated functions. It is thus an essential tool for modelling.

The **objectives of this course** are to familiarize the reader with the principal models (which are often partial differential equations), their methods of numerical solution and their optimization. Of course, the ambition of this course is to give a foundation which will allow future engineers either in a design department or in research and development to create **new models** and **new numerical algorithms** for more complicated problems not discussed here. However, even those not destined for such a career are interested in understanding the fundamentals of numerical simulation. Indeed, many industrial or political decisions will be taken from now on having faith in calculations or numerical simulations. It is thus essential that the decision-makers are capable of judging the **quality** and of the **reliability** of the calculations which are presented to them. This course will allow them to understand the first criteria which guarantee the validity and the relevance of numerical simulations.

The plan of this course is the following. After a first chapter of introduction to the principal ‘classical’ models and to their numerical solution, Chapter 2 is dedicated to the study of the numerical method of **finite differences**. These first two chapters allow us to go very quickly to some essential numerical questions which motivate the theoretical developments that follow. Chapters 3, 4, and 5 are dedicated to the theoretical solution by the **variational approach** of stationary (independent of time) models. They also give the foundation of a very important numerical method, called the **finite element** method, which is presented in detail in Chapter 6. The finite element method is the basis of many pieces of industrial or academic software. Chapters 7 and 8 discuss the solution of **nonstationary problems** (or of evolution in time), from both the theoretical and numerical points of view. If the first eight chapters are dedicated to numerical analysis, the last three treat **optimization**. Chapter 9 presents a series of concrete examples of optimization problems and gives a theory of existence of solutions to these problems. Chapter 10 derives the (necessary or sufficient) conditions for optimality of the solutions. These conditions are important as much from the theoretical as the numerical point of view. They allow us to characterize the optima, and they are the foundation of the numerical algorithms that we describe. Finally, chapter 11 is an introduction to **operational research**. After having studied linear programming, we give an outline of combinatorial optimization methods (that is to say optimization in discrete variables) which are essential for the optimal planning of resources and tasks in all large companies. Each chapter starts with an introduction which gives the plan and the principal ideas.

The length of this course should not worry the reader: the course contains numerous supplementary developments which allow the curious reader ‘to go a little further’ and to make the link with other works or other disciplines. It is, therefore, more a work of reference than the exact transcription of a lecture course.

To finish this introduction we give some practical information. As far as possible, this course is intended to be ‘self-contained’ to avoid frequent references to other works. This

is particularly sensible for many results from analysis which here are only useful, but not essential, technical tools. Statement without proof would amount to using them as ‘black boxes’ which gives them the flavour of an artificial ‘recipe’. As far as possible, we have therefore included their proof, but more as information and to ‘demystify’ than for the theoretical interest of the mathematical arguments. In order to distinguish them we use, for all these difficult passages or those of complementary interest, smaller characters like these. The reader should therefore consider these passages in small characters as ‘outside of the programme’. *The statements of results or of definitions are in italic characters like these.* The exercises are in sans serif characters like these. The end of a proof is indicated by the character  $\square$ , while the end of a remark or of an example is indicated by the character  $\bullet$ . An index is available at the end of the work.

The answers to exercises will be published in French. Most of the computer programs which implement the numerical methods studied, and which have allowed us to produce the figures in this work, are available on the website

[http://www.cmap.polytechnique.fr/~allaire/course\\_X\\_annee2.html](http://www.cmap.polytechnique.fr/~allaire/course_X_annee2.html)

where the reader can download them freely. The finite difference schemes, as well as the finite element method in one dimension, have been programmed in the language Scilab developed by INRIA and ENPC, available free on the website

<http://www.scilab.org>

while the results of the finite element method in two dimensions have been obtained with the help of the program FreeFem++ developed by F. Hecht and O. Pironneau and also available free on the website

<http://www.freefem.org>

In addition, most of the two-dimensional figures and all of the three-dimensional figures have been drawn with the help of graphical program xd3d developed by François Jouve at the École Polytechnique and also available free on the website

<http://www.cmap.polytechnique.fr/~jouve/xd3d>

Let us indicate another web address for the curious reader who wants to know more about the history of mathematics or the life of some mathematicians cited in this course

<http://www-history.mcs.st-and.ac.uk/history>

The reader who would like to keep up to date with the progress and advances of applied mathematics would benefit to consult the site of the Société de Mathématiques Appliquées et Industrielles

<http://smai.emath.fr>

or that of its American colleague, the Society for Industrial and Applied Mathematics

<http://www.siam.org>

The level of this course is introductory and it does not need any other prerequisites other than the level of knowledge gained in the first few years of university. We recognize that it is difficult to show much originality in this subject which is already classical in the literature. In particular, our course owes much to its predecessors and particularly to the course by B. Larrouturou, P.-L. Lions, and P.-A. Raviart from which it sometimes borrows heavily. The author thanks all those who have reread parts of the manuscript, particularly Frédéric Bonnans, Bruno Després and Bertrand Maury. A special mention is due to Stéphane Gaubert, who has co-written chapter 11, and also to Olivier Pantz, who has reread the entire manuscript with great care and who has checked the exercises and written



the corrections. The author thanks in advance all those who will point out the inevitable errors or imperfections of this edition, for example, by email to the address `gregoire.allaire@polytechnique.fr`

G. Allaire  
Paris, July 7, 2005

# 1 Introduction to mathematical modelling and numerical simulation

---

## 1.1 General introduction

This chapter is an introduction to two distinct, but closely linked, aspects of applied mathematics: **mathematical modelling** and **numerical simulation**. A mathematical model is a representation or an abstract interpretation of physical reality that is amenable to analysis and calculation. Numerical simulation allows us to calculate the solutions of these models on a computer, and therefore to simulate physical reality. In this book, the models we shall study will be partial differential equations (or PDEs), that is, differential equations in several variables (time and space, for example).

For the moment we shall put aside a third fundamental aspect of applied mathematics, that is, the mathematical analysis of models to which we shall return in a little more depth in later chapters. We need, in some way, to both motivate and justify this necessary intrusion of mathematical analysis. We shall see that the numerical calculation of the solutions of these physical models sometimes has some unpleasant surprises which can only be explained by a sound understanding of their mathematical properties. Once again we recall the fundamental multidisciplinary character of applied mathematics, and therefore of numerical simulation, which combines mathematics, computer science, and engineering.

Although most of the problems and applications which motivate applied mathematics are fundamentally **nonlinear** (see, for example, [12], [27]), we confine ourselves

in this work to linear problems for simplicity. Likewise, we only consider deterministic problems, that is, with no random or stochastic components. Finally, in order for this chapter to be introductory and easily accessible, we shall often be a little imprecise in our mathematical arguments. The more rigorous reader can be reassured that we shall return to the concepts introduced in this way more carefully in the following chapter.

The plan of this chapter is the following. Section 1.2 is devoted to an elementary example of modelling which leads to **the heat flow equation**. Section 1.3 is a quick review of the principal PDEs that we meet in the usual models in mechanics, physics, or engineering sciences. Section 1.4 is an informal introduction to numerical analysis and the **finite difference** method. Finally, in the Section 1.5 we give the definition of a **well-posed problem** as well as a (brief) classification of PDEs.

## 1.2 An example of modelling

Modelling represents a considerable part of the work of an applied mathematician and requires a thorough knowledge, not only of applied mathematics, but also of the scientific discipline to which it is applied. In fact, in many cases the mathematical model may not yet be established, or we must select the pertinent one from among several possibilities, or we must simplify known models which are too complex. However, in an introductory presentation of the discipline it is not possible to do justice to this step of the modelling process: we must begin by learning the basic properties of applied mathematics! This is why we limit ourselves to describing the derivation of a well-known classical physical model, and we refer the reader who wishes to know more to more specialised works.

The model which we shall describe is known as the **heat flow equation**, or the diffusion equation.

Let us consider a domain  $\Omega$  in  $N$  space dimensions (denoted by  $\mathbb{R}^N$ , with in general  $N = 1, 2$ , or  $3$ ) which we assume is occupied by a homogeneous, isotropic material which conducts heat. We denote the space variable by  $x$ , that is a point of  $\Omega$ , and the time variable by  $t$ . The heat sources in  $\Omega$  (possibly nonuniform in time and space) are represented by a given function  $f(x, t)$ , while the temperature is an unknown function  $\theta(x, t)$ . The quantity of the heat is proportional to the temperature  $\theta$  and is therefore  $c\theta$  where  $c$  is a physical constant (which depends on the material) called the specific heat. To calculate the temperature  $\theta$ , we write down the **law of conservation of energy** or of heat. In an elementary volume  $V$  contained in  $\Omega$ , the variation in time of the amount of heat is the balance of that produced by the sources and that which leaves or returns through the element boundaries. In other words,

$$\frac{d}{dt} \left( \int_V c\theta \, dx \right) = \int_V f \, dx - \int_{\partial V} q \cdot n \, ds, \quad (1.1)$$

where  $\partial V$  is the boundary of  $V$  (with surface element  $ds$ ),  $n$  is the outward unit

normal from  $V$ , and  $q$  is the heat flux vector. If we apply Gauss's theorem we obtain

$$\int_{\partial V} q \cdot n \, ds = \int_V \operatorname{div} q \, dx.$$

Gathering together the different terms in (1.1) and using the fact that the elementary volume  $V$  is independent of time, we deduce the energy conservation equation

$$c \frac{\partial \theta}{\partial t} + \operatorname{div} q = f \quad (1.2)$$

which holds at every point  $x \in \Omega$  and for all time  $t$ . We recall that the divergence operator is defined by

$$\operatorname{div} q = \sum_{i=1}^N \frac{\partial q_i}{\partial x_i} \text{ with } q = (q_1, \dots, q_N)^T.$$

We must now link the heat flow to the temperature, by what is called a **constitutive law**. In this case, we use Fourier's law which says that the heat flux is proportional to the temperature gradient

$$q = -k \nabla \theta T \quad (1.3)$$

where  $k$  is a positive constant (which depends on the material) called the thermal conductivity. Remember that the gradient operator is defined by

$$\nabla \theta = \left( \frac{\partial \theta}{\partial x_1}, \dots, \frac{\partial \theta}{\partial x_N} \right)^T.$$

By combining the conservation law (1.2) and the constitutive law (1.3), we obtain an equation for the temperature  $\theta$

$$c \frac{\partial \theta}{\partial t} - k \Delta \theta = f,$$

where  $\Delta = \operatorname{div} \nabla$  is the Laplacian operator given by

$$\Delta \theta = \sum_{i=1}^N \frac{\partial^2 \theta}{\partial x_i^2}.$$

This equation is valid in the entire domain  $\Omega$  and we must add another relation, called a **boundary condition**, which describes what happens at the boundary  $\partial \Omega$  of the domain, and another relation which describes the initial state of the temperature. By convention, we choose the instant  $t = 0$  to be the initial time, and we impose an **initial condition**

$$\theta(t = 0, x) = \theta_0(x), \quad (1.4)$$

where  $\theta_0$  is the function giving the initial distribution of the temperature in the domain  $\Omega$ . The type of boundary condition depends on the physical context. If the domain is

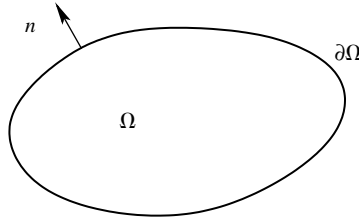


Figure 1.1. Unit normal vector oriented to the exterior.

surrounded by a region of constant temperature, then, by rescaling the temperature, the temperature satisfies the Dirichlet boundary condition

$$\theta(t, x) = 0 \quad \text{for all } x \in \partial\Omega \text{ and } t > 0. \quad (1.5)$$

If the domain is assumed to be adiabatic or thermally isolated from the exterior, then the heat flux across the boundary is zero and the temperature satisfies the Neumann boundary condition

$$\frac{\partial\theta}{\partial n}(t, x) \equiv n(x) \cdot \nabla\theta(t, x) = 0 \quad \text{for all } x \in \partial\Omega \text{ and } t > 0, \quad (1.6)$$

where  $n$  is the unit outward normal to  $\Omega$  (see Figure 1.1). An intermediate situation can happen: the heat flux across the boundary is proportional to the jump in temperature from the exterior to the interior, and the temperature satisfies the Fourier (or Robin) boundary condition

$$\frac{\partial\theta}{\partial n}(t, x) + \alpha\theta(t, x) = 0 \quad \text{for all } x \in \partial\Omega, \text{ and } t > 0 \quad (1.7)$$

where  $\alpha$  is a positive constant. As we must choose a boundary condition (as one of the steps in the modelling), we shall take the Dirichlet boundary condition (1.5). Finally, gathering together the equation, the initial value, and the boundary condition satisfied by the temperature, we obtain the heat equation

$$\begin{cases} c \frac{\partial\theta}{\partial t} - k\Delta\theta = f & \text{for } (x, t) \in \Omega \times \mathbb{R}_*^+ \\ \theta(t, x) = 0 & \text{for } (x, t) \in \partial\Omega \times \mathbb{R}_*^+ \\ \theta(t = 0, x) = \theta_0(x) & \text{for } x \in \Omega \end{cases} \quad (1.8)$$

Problem (1.8) therefore comprises a PDE equipped with boundary conditions and an initial value. Because of the boundary conditions, we say that (1.8) is a **boundary value problem**, but we also say that it is a **Cauchy problem** because of the initial value.

**Remark 1.2.1** In this model of heat propagation we must make the physical units precise: the temperature  $\theta$  is expressed in degrees Kelvin ( $K$ ), the specific heat  $c$  in Joules per kilogram per degree Kelvin ( $J/(kg \times K)$ ), the thermal conductivity (per unit of mass)  $k$  in Joules metre squared per kilogramme per degree Kelvin per second ( $Jm^2/(kg \times K \times s)$ ). From a mathematical point of view, we shall frequently neglect these units, and also assume that the constants  $c$  and  $k$  are equal to 1 (this is equivalent to making the physical quantities nondimensional). •

**Remark 1.2.2** We have mentioned three types of boundary condition, Dirichlet, Neumann, and Fourier (but there are others) which hold on the entire boundary  $\partial\Omega$ . Of course, we can easily imagine situations where the boundary conditions are mixed: Dirichlet on  $\partial\Omega_D$ , Neumann on  $\partial\Omega_N$ , and Fourier on  $\partial\Omega_F$ , with  $\partial\Omega_D, \partial\Omega_N, \partial\Omega_F$  being a partition of the boundary  $\partial\Omega$ . •

**Remark 1.2.3** The heat flow equation (1.8) is **linear** in the sense that its solution  $\theta$  depends linearly on the data  $(f, \theta_0)$ . In physics, this property is often described in terms of a superposition principle: a linear combination of data  $(f, \theta_0)$  leads to a solution  $\theta$  which is the same linear combination of solutions corresponding to each term of the decomposition of data. From a physical point of view, linearity is only one hypothesis among many. Indeed, for problems with a strong variation in temperature, Fourier's law is false, and it should be corrected by assuming that the thermal conductivity  $k$  depends on the temperature  $\theta$  and its gradient  $\nabla\theta$  (which makes the problem nonlinear). Even worse, for very rapid phenomena (explosions, for example) it is necessary to abandon the assumption of the proportionality of the heat flux  $q$  to the temperature gradient  $\nabla\theta$ . Indeed, this hypothesis (which initially appears 'natural') leads to the following paradox: the heat propagates with infinite velocity in the domain  $\Omega$ . We shall see later (see Remark 1.2.9) how to reach this paradox. Let us remember for the moment that modelling is making hypotheses and describing their domain of validity. •

**Remark 1.2.4** Problem (1.8) is not just a model of heat propagation. In fact it has a universal character, and we find it in many unrelated phenomena (we simply change the names of the variables). For example, (1.8) is also known as **the diffusion equation**, and models the diffusion or migration of a density or concentration across the domain  $\Omega$  (imagine a pollutant diffusing in the atmosphere, or a chemical species migrating in a substrate). In this case,  $\theta$  is the concentration or the density in question,  $q$  is the mass flux,  $k$  is the diffusivity, and  $c$  is the volume density of the species. Likewise, the conservation law (1.2) is a mass balance, while the constitutive law (1.3) is called Fick's law. •

**Remark 1.2.5** Problem (1.8) also occurs in finance where it is called the **Black–Scholes model**. A variant of (1.8) allows us to find the value of an option to buy (or call option) a stock, which is initially worth  $x$ , for price  $k$  at some time in the future  $T$ .

This value is the solution  $u$  of

$$\begin{cases} \frac{\partial u}{\partial t} - ru + 1/2rx \frac{\partial u}{\partial x} + 1/2\sigma^2 x^2 \frac{\partial^2 u}{\partial x^2} = 0 & \text{for } (x, t) \in \mathbb{R} \times (0, T) \\ u(t = T, x) = \max(x - k, 0) & \text{for } x \in \mathbb{R} \end{cases} \quad (1.9)$$

More precisely,  $u(0, x)$  is the value at time  $t = 0$  of the call option with exercise price  $k$  at the exercise time  $T > 0$ , and with value  $x$  at  $t = 0$ . The volatility is denoted by  $\sigma$  and the interest rate by  $r$ . We remark that (1.9) is a final value and not an initial value problem, but that the sign of the second space derivative is opposite to that in (1.8). Consequently, after reversing the time, (1.9) is a parabolic equation. •

Numerous variants of the heat equation (1.8) exist, some of which we shall now explore. Up until now we have assumed that heat propagates in a fixed medium, or at least a still medium. Let us now assume that it propagates in a moving medium, for example, a fluid moving with velocity  $V(x, t)$  (a vector valued function in  $\mathbb{R}^N$ ). Then, we must now change the constitutive law since the heat flux is the sum of a diffusive flux (as before) and a convective flux (proportional to the velocity  $V$ ), and proceeding similarly to the arguments above leads us to the **convection–diffusion** problem

$$\begin{cases} c \frac{\partial \theta}{\partial t} + cV \cdot \nabla \theta - k \Delta \theta = f & \text{in } \Omega \times \mathbb{R}_*^+ \\ \theta = 0 & \text{on } \partial\Omega \times \mathbb{R}_*^+ \\ \theta(t = 0, x) = \theta_0(x) & \text{in } \Omega \end{cases} \quad (1.10)$$

The difference between (1.8) and (1.10) is the appearance of a convection term. We measure the balance between this new convection term and the diffusion term by a dimensionless number called the **Péclet number**, defined by

$$\text{Pe} = \frac{cVL}{k}, \quad (1.11)$$

where  $L$  is a characteristic length of the problem (for example, the diameter of the domain  $\Omega$ ). If the Péclet number is very small then the diffusive effects dominate the convective effects, and model (1.8) is sufficient to describe the phenomenon. If the Péclet number is neither small nor large (we say that it is the order of unity), then model (1.10) is more realistic than (1.8). On the other hand, if the Péclet number is very large, we can simplify (1.10) by removing the diffusion term. We then obtain the equation known as the **advection** equation

$$\begin{cases} c \frac{\partial \theta}{\partial t} + cV \cdot \nabla \theta = f & \text{in } \Omega \times \mathbb{R}_*^+ \\ \theta(t, x) = 0 & \text{for } (x, t) \in \partial\Omega \times \mathbb{R}_*^+ \text{ if } V(x) \cdot n(x) < 0 \\ \theta(t = 0, x) = \theta_0(x) & \text{in } \Omega \end{cases} \quad (1.12)$$

We note the difference in the boundary condition of (1.12) with respect to that of (1.10): we no longer impose that the temperature  $\theta$  is zero everywhere on the boundary  $\partial\Omega$  but only on those parts of the boundary where the velocity  $V$  is re-entrant.

We have therefore described three models of heat propagation by convection and diffusion, (1.8), (1.10), (1.12), which have different regimes of validity depending on different values of the Péclet number. Of course, the analytical or numerical solution of these three problems is very different. This is reflected in the current state of mathematical modelling: there are several competing models and we must choose the ‘best’.

In order to understand better the fundamental differences which exist between these models, we temporarily restrict ourselves to the case where  $\Omega = \mathbb{R}$  the whole real line (which rids us of the question of the boundary conditions), where the source term  $f$  is zero, and where the velocity  $V$  is constant. We can then explicitly calculate solutions of these models. For example, (1.10) becomes

$$\begin{cases} \frac{\partial \theta}{\partial t} + V \frac{\partial \theta}{\partial x} - \nu \frac{\partial^2 \theta}{\partial x^2} = 0 & \text{for } (x, t) \in \mathbb{R} \times \mathbb{R}_*^+ \\ \theta(t = 0, x) = \theta_0(x) & \text{for } x \in \mathbb{R} \end{cases} \quad (1.13)$$

with  $\nu = k/c$ , which has solution

$$\theta(t, x) = \frac{1}{\sqrt{4\pi\nu t}} \int_{-\infty}^{+\infty} \theta_0(y) \exp\left(-\frac{(x - Vt - y)^2}{4\nu t}\right) dy. \quad (1.14)$$

A solution of (1.8) is easily obtained by setting  $V = 0$  in the expression (1.14).

**Exercise 1.2.1** We assume that the initial condition  $\theta_0$  is continuous and uniformly bounded in  $\mathbb{R}$ . Verify that (1.14) is a solution of (1.13).

With the same simplifying hypotheses, the advection equation becomes

$$\begin{cases} \frac{\partial \theta}{\partial t} + V \frac{\partial \theta}{\partial x} = 0 & \text{for } (x, t) \in \mathbb{R} \times \mathbb{R}_*^+ \\ \theta(t = 0, x) = \theta_0(x) & \text{for } x \in \mathbb{R} \end{cases} \quad (1.15)$$

We verify that

$$\theta(t, x) = \theta_0(x - Vt) \quad (1.16)$$

is a solution of the equation (1.15).

**Exercise 1.2.2** We assume that the initial data  $\theta_0$  is differentiable and uniformly bounded over  $\mathbb{R}$ . Verify that (1.16) is a solution of (1.15). Show that (1.16) is the limit of (1.14) as the parameter  $\nu$  tends to zero.

**Remark 1.2.6** If we solve the heat flow equation (1.8) on a bounded interval (and not in the whole space), we can also calculate an explicit solution by using Fourier analysis (see [4], [38]). This solution would be a little less ‘explicit’ than (1.14) as it is defined as the sum of an infinite series. We remark that it was precisely to solve the heat flow equation that Fourier invented the analysis which takes his name. •



**Remark 1.2.7** The role of time is fundamentally different in equations (1.8) and (1.12). Indeed, assuming that the source term is zero,  $f = 0$ , if we change the sign of time  $t$  and that of the velocity, the advection equation (1.12) is unchanged (when we change the time we change the current). Conversely, a change in the sign of time in the heat flow equation (1.8) cannot be compensated by any variation in the sign of the data. This is obvious in the explicit form of the solution: (1.16) is invariant by changing the sign of  $t$  and  $V$ , whereas (1.14) (with  $V = 0$ ) decreases in time, indicating the ‘arrow’ of time. We say that the advection equation is **reversible** in time, while the heat flow equation is **irreversible** in time. This mathematical observation is confirmed by physical intuition: some phenomena are reversible in time, others are not (like the diffusion of a drop of milk in a cup of tea). •

**Remark 1.2.8** Another fundamental difference between equations (1.8) and (1.12) lies with the property of **invariance with respect to change of scale**. Let us assume that the source term is zero,  $f = 0$ . It is easy to see that if  $\theta(x, t)$  is a solution of the heat flow equation (1.8), then, for all  $\lambda > 0$ ,  $\theta(x/\lambda, t/\lambda^2)$  is also a solution of the same equation (for a different initial value). Likewise, assuming that the velocity  $V$  is constant, if  $\theta(x, t)$  is a solution of the advection equation (1.12), then  $\theta(x/\lambda, t/\lambda)$  is also a solution. We see that the scaling of time is not the same in both cases. We also remark that, in both cases, the equations are invariant under translation in space and in time. •

**Remark 1.2.9** A surprising property (from the physical point of view) of the heat flow equation (1.8) is that the solution in  $(x, t)$  depends on all the initial values in  $\mathbb{R}$  (see formula (1.14)). In particular, in the case of (1.13), if the initial data is positive with compact support, then for all time  $t > 0$  (no matter how small) the solution is strictly positive over all  $\mathbb{R}$ : in other words, the heat propagates ‘instantaneously’ to infinity. We say that the heat **propagates with an infinite velocity** (which is clearly a limitation of the model). On the other hand, in the advection equation (1.15) the initial data is convected with velocity  $V$  (see formula (1.16)): therefore there is a **finite velocity of propagation**. •

**Remark 1.2.10** Thanks to the explicit formulas (1.14) and (1.16), we easily verify that the solutions of the convection–diffusion equation (1.13) and of the advection equation (1.15) satisfy the property

$$\min_{x \in \mathbb{R}} \theta_0(x) \leq \theta(x, t) \leq \max_{x \in \mathbb{R}} \theta_0(x) \quad \text{for all } (x, t) \in \mathbb{R} \times \mathbb{R}^+,$$

which is called the **maximum principle**. This property (which is equally important from the point of view of both mathematics and physics) extends to more general forms of the convection–diffusion equation (1.10) and of the advection equation (1.12). We shall study it more carefully later. •

## 1.3 Some classical models

In this section we shall quickly describe some classical models. Our goal is to present the principal classes of PDEs which we shall study later, and to show that these equations play a very important role in diverse scientific areas. From now on, we shall nondimensionalize all the variables, which will allow us to set the constants in the models equal to 1.

### 1.3.1 The heat flow equation

As we have seen, the heat flow equation appears as a model in many problems in science and engineering. It is written

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{in } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{on } \partial\Omega \times \mathbb{R}_*^+ \\ u(t = 0) = u_0 & \text{in } \Omega. \end{cases} \quad (1.17)$$

This equation is first order in time and second order in space (the order is that of the highest partial derivatives). We shall say that this equation is parabolic (see Section 1.5.2). We have already seen some properties of this equation: irreversibility in time, propagation with infinite velocity, and the maximum principle.

**Exercise 1.3.1** We shall find a property of exponential decrease in time (see formula (1.14)) of the solution of the heat flow equation (1.17) in a bounded domain. In one space dimension, we set  $\Omega = (0, 1)$  and we assume that  $f = 0$ . Let  $u(t, x)$  be a regular solution of (1.17). Multiplying the equation by  $u$  and integrating with respect to  $x$ , establish the equality

$$\frac{1}{2} \frac{d}{dt} \left( \int_0^1 u^2(t, x) dx \right) = - \int_0^1 \left| \frac{\partial u}{\partial x}(t, x) \right|^2 dx.$$

Show that every continuously differentiable function  $v(x)$  on  $[0, 1]$ , such that  $v(0) = 0$ , satisfies the Poincaré inequality

$$\int_0^1 v^2(x) dx \leq \int_0^1 \left| \frac{dv}{dx}(x) \right|^2 dx.$$

From this, deduce the exponential decrease in time of  $\int_0^1 u^2(t, x) dx$ .

### 1.3.2 The wave equation

The wave equation models propagation of waves or vibration. For example, in two space dimensions it is a model to study the vibration of a stretched elastic membrane (like the skin of a drum). In one space dimension, it is also called the vibrating cord equation. At rest, the membrane occupies a plane domain  $\Omega$ . Under the action of

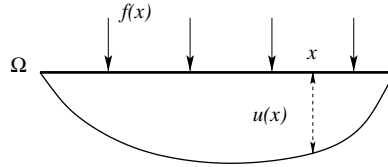


Figure 1.2. Displacement of an elastic cord.

a force normal to the plane with intensity  $f$ , it deforms and its normal displacement is denoted by  $u$  (see Figure 1.2). We assume that it is fixed at the boundary, which gives a Dirichlet boundary condition. The wave equation with solution  $u$  is given by

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta u = f & \text{in } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{on } \partial\Omega \times \mathbb{R}_*^+ \\ u(t = 0) = u_0 & \text{in } \Omega \\ \frac{\partial u}{\partial t}(t = 0) = u_1 & \text{in } \Omega \end{cases} \quad (1.18)$$

We note that this equation is second order in time and that we therefore need two initial conditions for  $u$ . We say that this equation is hyperbolic (see Section 1.5.2).

**Exercise 1.3.2** We work in  $N = 1$  space dimensions. We assume that the initial data  $u_0$  and  $u_1$  are regular functions, and that  $f = 0$  with  $\Omega = \mathbb{R}$ . We note that  $U_1$  is a primitive of  $u_1$ . Verify that

$$u(t, x) = \frac{1}{2} (u_0(x+t) + u_0(x-t)) + \frac{1}{2} (U_1(x+t) - U_1(x-t)), \quad (1.19)$$

is the unique solution of (1.18) in the class of regular functions.

The wave equation shares, with the advection equation (1.12), the important property of **propagation with finite velocity**. Indeed, exercise 1.3.3 shows that the solution at a point  $(x, t)$  does not depend on all the initial data but only on the values in a restricted interval called the **domain of dependence** (or light cone; see Figure 1.3). We recall that this property is not shared by the heat flow equation since it is clear, from formula (1.14), that the solution in  $(x, t)$  depends on all the values of the initial data.

Another property of the wave equation is its invariance under the change of direction of time. If we change  $t$  to  $-t$ , the form of the equation does not change. We can therefore ‘integrate’ the wave equation in the positive or negative time directions in the same way. We say that the wave equation is **reversible in time**.

**Exercise 1.3.3** Verify that the solution (1.19) at the point  $(x, t)$  only depends on the values of the initial data  $u_0$  and  $u_1$  in the segment  $[x - t, x + t]$ . Verify also that  $u(-t, x)$  is a solution of (1.18) in  $\Omega \times \mathbb{R}_*^+$  if we change the sign of the initial velocity  $u_1(x)$ .

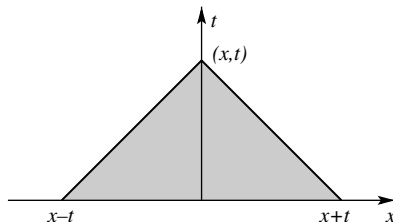


Figure 1.3. Domain or cone of dependence of the wave equation.

**Exercise 1.3.4** We propose showing a principle of conservation of energy for the wave equation (1.18) without using the explicit formula (1.19). In one space dimension, we set  $\Omega = (0, 1)$  and we assume  $f = 0$ . Let  $u(t, x)$  be a regular solution of (1.18). Multiplying the equation by  $\partial u / \partial t$  and integrating with respect to  $x$ , establish the energy equality

$$\frac{d}{dt} \left( \int_0^1 \left| \frac{\partial u}{\partial t}(t, x) \right|^2 dx + \int_0^1 \left| \frac{\partial u}{\partial x}(t, x) \right|^2 dx \right) = 0.$$

Compare this with what happens for the heat equation.

### 1.3.3 The Laplacian

For certain choices of source term  $f$ , the solution of the heat flow equation (1.17) reaches a **steady** (or stationary) state, that is,  $u(t, x)$  tends to a limit  $u_\infty(x)$  as time  $t$  tends to infinity. Often, it is interesting to calculate this steady state directly. In this case, for a source term  $f(x)$  which is independent of time, we solve an equation which is second order in space

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (1.20)$$

which we call the Laplacian or Laplace's equation. We say that this equation is elliptic (see Section 1.5.2). We remark that the Laplacian is also the stationary version of the wave equation (1.19). The Laplacian also occurs in numerous fields of science and engineering. For example, (1.20) models the vertical displacement of an elastic membrane subjected to a normal force  $f$  and fixed around its boundary.

### 1.3.4 Schrödinger's equation

Schrödinger's equation describes the evolution of the wave function  $u$  of a particle subject to a potential  $V$ . Recall that  $u(t, x)$  is a function of  $\mathbb{R}^+ \times \mathbb{R}^N$  with values in  $\mathbb{C}$  and that the square of its modulus  $|u|^2$  is interpreted as the probability that the particle is found at the point  $(t, x)$ . The potential  $V(x)$  is a real-valued function. The wave function is a solution of

$$\begin{cases} i \frac{\partial u}{\partial t} + \Delta u - Vu = 0 & \text{in } \mathbb{R}^N \times \mathbb{R}_*^+ \\ u(t=0) = u_0 & \text{in } \mathbb{R}^N \end{cases} \quad (1.21)$$

There are no boundary conditions in (1.21) since the equation holds over the whole of space (which has no boundary). Nevertheless, we shall see that a 'reasonable' choice of function space in which to look for the solution implies *de facto* a condition of decay to infinity of  $u$  which can be interpreted as a boundary condition at infinity.

**Exercise 1.3.5** We propose to show principles of energy conservation for Schrödinger's equation (1.21). Let  $u(t, x)$  be a regular solution of (1.21) in one space dimension which decreases to zero (as does  $\partial u / \partial x$ ) as  $|x| \rightarrow +\infty$ . Show that for every differentiable function  $v(t)$  we have

$$\mathcal{R} \left( \frac{\partial v}{\partial t} \bar{v} \right) = \frac{1}{2} \frac{\partial |v|^2}{\partial t},$$

where  $\mathcal{R}$  denotes the real part and  $\bar{v}$  the complex conjugate of  $v$ . Multiplying the equation by  $\bar{u}$  and integrating with respect to  $x$ , establish the energy equality

$$\int_{\mathbb{R}} |u(t, x)|^2 dx = \int_{\mathbb{R}} |u_0(x)|^2 dx.$$

Multiplying the equation by  $\partial \bar{u} / \partial t$ , show that

$$\int_{\mathbb{R}} \left( \left| \frac{\partial u}{\partial x}(t, x) \right|^2 + V(x) |u(t, x)|^2 \right) dx = \int_{\mathbb{R}} \left( \left| \frac{\partial u_0}{\partial x}(x) \right|^2 + V(x) |u_0(x)|^2 \right) dx.$$

### 1.3.5 The Lamé system

The Lamé system is a particular case of the linearized stationary elasticity equations which model deformations of a solid under the assumption of small deformations and of small displacements (see Section 5.3.1 for further details on the modelling). To obtain the Lamé system, we assume that the solid is homogeneous and isotropic and that it is fixed at the boundary. The principal difference from the preceding models is that here we have a **system** of equations, that is, several coupled equations. The solid at rest occupies the domain  $\Omega$  of the space  $\mathbb{R}^N$ . Under the action of a force  $f$

it deforms, and each point  $x$  moves to  $x + u(x)$ . The force  $f(x)$  is a vector-valued function of  $\Omega$  in  $\mathbb{R}^N$ , as is the displacement  $u(x)$ . This is a solution of

$$\begin{cases} -\mu\Delta u - (\mu + \lambda)\nabla(\operatorname{div} u) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (1.22)$$

where  $\lambda$  and  $\mu$  are two constants, the Lamé constants, which are characteristics of the homogeneous, isotropic material which comprises the solid. For mechanical reasons these constants satisfy  $\mu > 0$  and  $2\mu + N\lambda > 0$ . The Dirichlet boundary condition for  $u$  reflects the fact that the solid is assumed fixed and immovable at its boundary.  $\partial\Omega$ .

The system (1.22) has been written in vector notation. If we denote by  $f_i$  and  $u_i$ , for  $1 \leq i \leq N$ , the components of  $f$  and  $u$  in the canonical basis of  $\mathbb{R}^N$ , (1.22) is equivalent to

$$\begin{cases} -\mu\Delta u_i - (\mu + \lambda)\frac{\partial(\operatorname{div} u)}{\partial x_i} = f_i & \text{in } \Omega \\ u_i = 0 & \text{on } \partial\Omega \end{cases}$$

for  $1 \leq i \leq N$ . We remark that, if  $(\mu + \lambda) \neq 0$ , then the equations for each component  $u_i$  are coupled by the divergence term. Obviously, in  $N = 1$  dimension, the Lamé system has only one equation and reduces to the Laplacian.

### 1.3.6 The Stokes system

The Stokes system models the flow of a viscous incompressible fluid with small velocity. We assume that the fluid occupies a domain  $\Omega$  and that it adheres to the boundary, that is, its velocity is zero at the boundary (which leads to a Dirichlet boundary condition). Under the action of a force  $f(x)$  (a function of  $\Omega$  in  $\mathbb{R}^N$ ), the velocity  $u(x)$  (a vector) and the pressure  $p(x)$  (a scalar) are solutions of

$$\begin{cases} \nabla p - \mu\Delta u = f & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (1.23)$$

where  $\mu > 0$  is the fluid viscosity. We note that there are a further  $N$  equations  $\nabla p - \mu\Delta u = f$  (corresponding to the **conservation of momentum**), and one other equation  $\operatorname{div} u = 0$  called the **incompressibility condition** (which corresponds to **conservation of mass**). If the space dimension is  $N = 1$ , the Stokes system is uninteresting as we easily see that the velocity is zero and the pressure is a primitive of the force. On the other hand, in dimensions  $N \geq 2$ , the Stokes system makes good sense: in particular, there exist nontrivial incompressible velocity fields (take, for example, a curl).

### 1.3.7 The plate equations

We consider small elastic deformations of a thin plane plate (which is negligible in its other dimensions). If we denote by  $\Omega$  the average surface of the plate, and  $f(x)$

(a function of  $\Omega$  in  $\mathbb{R}$ ) the resultant normal of the forces, then the normal component of the displacement  $u(x)$  (a scalar) is the solution of the thin plate equation

$$\begin{cases} \Delta(\Delta u) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega \end{cases} \quad (1.24)$$

where we denote by  $\frac{\partial u}{\partial n} = \nabla u \cdot n$  with  $n$  the outward unit normal vector to  $\partial\Omega$ . We remark that this is a partial differential equation which is fourth order in space (also called the bi-Laplacian). This is why it is necessary to have two boundary conditions. These boundary conditions represent the clamping of the plate (there is neither displacement nor rotation of the edge of the plate).

We remark that it is possible to justify the plate equation (1.24) asymptotically from the Lamé system (1.22) by letting the thickness of the plate to tend to zero. This is an example of mathematical modelling.

## 1.4 Numerical calculation by finite differences

### 1.4.1 Principles of the method

Apart from some very particular cases, it is impossible to calculate explicitly the solutions of the different models presented above. It is therefore necessary to have recourse to numerical calculation on a computer to estimate these solutions both qualitatively and quantitatively. The principle of all methods for the numerical solution of PDEs is to obtain discrete numerical values (that is, a finite number) which ‘**approximate**’ (in a suitable sense, to be made precise) the exact solution. In this process we must be aware of two fundamental points: first, we do not calculate exact solutions but approximate ones; second, we **discretize** the problem by representing functions by a finite number of values, that is, **we move from the ‘continuous’ to the ‘discrete’**.

There are numerous methods for the numerical approximation of PDEs. We present one of the oldest and simplest, called the finite difference method (later we shall see another method, called the finite element method). For simplicity, we limit ourselves to one space dimension (see Section 2.2.6 for higher dimensions). For the moment, we shall only consider the practical principles of this method, that is, the construction of what we call the **numerical schemes**. We reserve the theoretical justification of these schemes for Chapter 2, that is, the study of their convergence (in what way the approximate discrete solutions are close to the exact continuous solutions).

To discretise the spatio-temporal continuum, we introduce a **space step**  $\Delta x > 0$  and a **time step**  $\Delta t > 0$  which will be the smallest scales represented by the numerical method. We define a mesh or discrete coordinates in space and time (see Figure 1.4)

$$(t_n, x_j) = (n\Delta t, j\Delta x) \quad \text{for } n \geq 0, \quad j \in \mathbb{Z}.$$

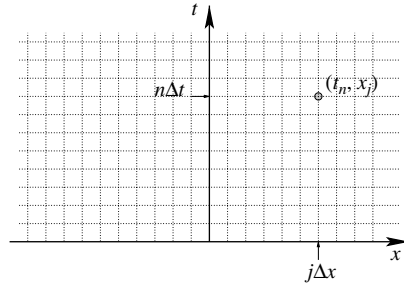


Figure 1.4. Finite difference mesh.

We denote by  $u_j^n$  the value of the discrete solution at  $(t_n, x_j)$ , and  $u(t, x)$  the (unknown) exact solution. The principle of the finite difference method is to replace the derivatives by finite differences by using Taylor series in which we neglect the remainders. For example, we approximate the second space derivative (the Laplacian in one dimension) by

$$-\frac{\partial^2 u}{\partial x^2}(t_n, x_j) \approx \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} \quad (1.25)$$

where we recall the Taylor formula

$$\begin{aligned} -u(t, x - \Delta x) + 2u(t, x) - u(t, x + \Delta x) = & -(\Delta x)^2 \frac{\partial^2 u}{\partial x^2}(t, x) \\ & - \frac{(\Delta x)^4}{12} \frac{\partial^4 u}{\partial x^4}(t, x) + \mathcal{O}((\Delta x)^6) \end{aligned} \quad (1.26)$$

If  $\Delta x$  is ‘small’, formula (1.25) is a ‘good’ approximation (it is natural, but not unique). The formula (1.25) is called **centred** since it is symmetric in  $j$ .

To discretize the convection–diffusion equation

$$\frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} - \nu \frac{\partial^2 u}{\partial x^2} = 0 \quad (1.27)$$

we must also discretize the convection term. A centred formula gives

$$V \frac{\partial u}{\partial x}(t_n, x_j) \approx V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x}$$

It only remains to do the same thing for the time derivative. Again we have a choice between finite difference schemes: centred or one sided. Let us look at three ‘natural’ formulas.



1. As a first choice, the centred finite difference

$$\frac{\partial u}{\partial t}(t_n, x_j) \approx \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t}$$

leads to a scheme which is completely symmetric with respect to  $n$  and  $j$  (called the centred scheme or **Richardson's scheme**)

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \quad (1.28)$$

Even though it is 'natural' **this scheme cannot calculate approximate solutions** of the convection–diffusion equation (1.27) (see the numerical example of Figure 1.5)! We shall justify the inability of this scheme to approximate the exact solution in Lemma 2.2.23. For the moment, we shall simply say that the difficulty comes from the centred character of the finite difference which approximates the time derivative.

2. A second choice is the one-sided upwind scheme (we go back in time) which gives the **backward Euler scheme**

$$\frac{\partial u}{\partial t}(t_n, x_j) \approx \frac{u_j^n - u_j^{n-1}}{\Delta t}$$

which leads to

$$\frac{u_j^n - u_j^{n-1}}{\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \quad (1.29)$$

3. The third choice is the opposite of the preceding: the downwind one-sided finite difference (we go forward in time; we also talk of the **forward Euler scheme**)

$$\frac{\partial u}{\partial t}(t_n, x_j) \approx \frac{u_j^{n+1} - u_j^n}{\Delta t}$$

which leads to

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \quad (1.30)$$

The principal difference between these last two schemes is that (1.29) is called **implicit** since we must solve a system of linear equations to calculate the values  $(u_j^n)_{j \in \mathbb{Z}}$  as functions of the preceding values  $(u_j^{n-1})_{j \in \mathbb{Z}}$ , while (1.30) is called **explicit** since it immediately gives the values  $(u_j^{n+1})_{j \in \mathbb{Z}}$  as a function of  $(u_j^n)_{j \in \mathbb{Z}}$ . The shift of 1 in the index  $n$  between the schemes (1.29) and (1.30) is only evident when we rewrite (1.30) in the form

$$\frac{u_j^n - u_j^{n-1}}{\Delta t} + V \frac{u_{j+1}^{n-1} - u_{j-1}^{n-1}}{2\Delta x} + \nu \frac{-u_{j-1}^{n-1} + 2u_j^{n-1} - u_{j+1}^{n-1}}{(\Delta x)^2} = 0.$$

In the three schemes which we have defined, there must be initial data to start the iterations in  $n$ : the initial values  $(u_j^0)_{j \in \mathbb{Z}}$  are defined by, for example,  $u_j^0 = u_0(j\Delta x)$  where  $u_0$  is the initial data of the convection–diffusion equation (1.27). We remark that the ‘bad’ centred scheme (1.28) has an additional difficulty in starting: for  $n = 1$  we also have to know the values  $(u_j^1)_{j \in \mathbb{Z}}$  which, therefore, must be calculated in another way (for example, by applying one of the two other schemes).

### 1.4.2 Numerical results for the heat flow equation

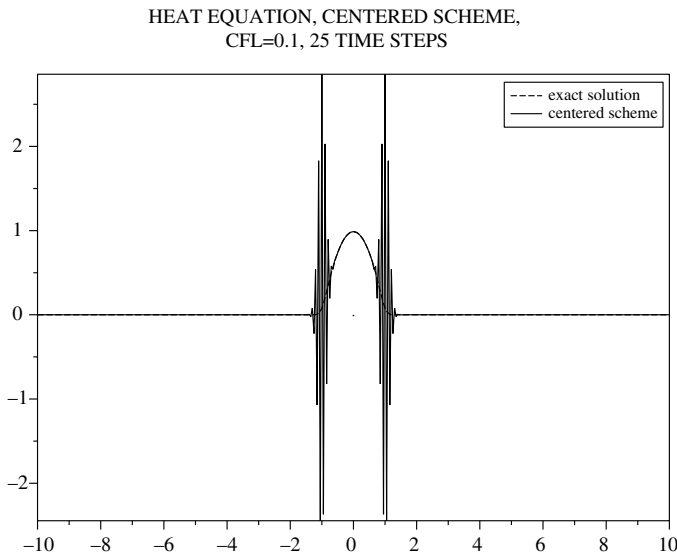


Figure 1.5. Unstable centred scheme with  $\nu\Delta t = 0.1(\Delta x)^2$ .

We start by making some simple numerical tests in the case where  $V = 0$  and  $\nu = 1$ , that is, **we solve the heat flow equation numerically**. As initial condition, we choose the function

$$u_0(x) = \max(1 - x^2, 0).$$

To be able to compare the numerical solutions with the exact (1.14), we want to work on the infinite domain  $\Omega = \mathbb{R}$ , that is, calculate, for each  $n \geq 0$ , an infinite number of values  $(u_j^n)_{j \in \mathbb{Z}}$ , but the computer will not allow this as the memory is finite! To a first approximation, we therefore replace  $\mathbb{R}$  by the ‘large’ domain  $\Omega = (-10, +10)$  equipped with Dirichlet boundary conditions. The validity of this approximation is confirmed by the numerical calculations below. We fix the space step at  $\Delta x = 0.05$ : there are therefore 401 values  $(u_j^n)_{-200 \leq j \leq +200}$  to calculate. We should remember that the values  $u_j^n$  calculated by the computer are subject to rounding errors and are therefore not the exact values of the difference scheme: nevertheless, in the calculations presented here, these rounding errors are completely negligible and are in no way

responsible for the phenomena which we shall observe. On all the figures we show the exact solution, calculated by the explicit formula (1.14), and the approximate numerical solution under consideration.

Let us first look at the outcome of the centred scheme (1.28): since as we have said, this scheme is not able to calculate approximate solutions of the heat flow equation. Whatever the choice of the time step  $\Delta t$ , this scheme is **unstable**, that is the numerical solution oscillates unboundedly if we decrease the step sizes  $\Delta x$  and  $\Delta t$ . This highly characteristic phenomenon (which appears rapidly) is illustrated by Figure 1.5. We emphasize that **whatever the choice** of steps  $\Delta t$  and  $\Delta x$ , we see these oscillations (which are nonphysical). We say that the scheme is unconditionally unstable. A rigorous justification will be given in the following chapter (see lemma 2.2.23).

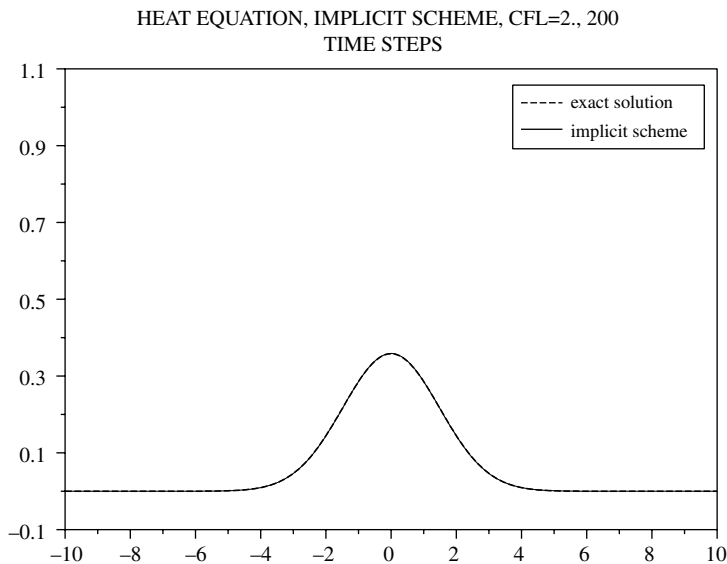


Figure 1.6. Implicit scheme with  $\nu\Delta t = 2(\Delta x)^2$ .

Contrary to the preceding scheme, the implicit scheme (1.29) calculates ‘good’ approximate solutions of the heat flow equation **whatever** the time step  $\Delta t$  (see Figure 1.6). In particular, we never see numerical oscillation for any choice of steps  $\Delta t$  and  $\Delta x$ . We say that the implicit scheme is unconditionally stable.

Let us now consider the explicit scheme (1.30): numerical experiments show that we obtain numerical oscillations depending on the time step  $\Delta t$  (see Figure 1.7). The stability limit is easy to find experimentally: if the choice of steps  $\Delta t$  and  $\Delta x$  **satisfy** the condition

$$2\nu\Delta t \leq (\Delta x)^2 \tag{1.31}$$

the scheme is stable, while if (1.31) is not satisfied, then the scheme is unstable. We say that the explicit scheme is conditionally stable. The stability condition (1.31)

is **one of the simplest but most profound observations in numerical analysis**. It was discovered in 1928 (before the appearance of the first computers!) by Courant, Friedrichs, and Lewy. It takes the name **CFL condition or the Courant, Friedrichs, Lewy condition**.

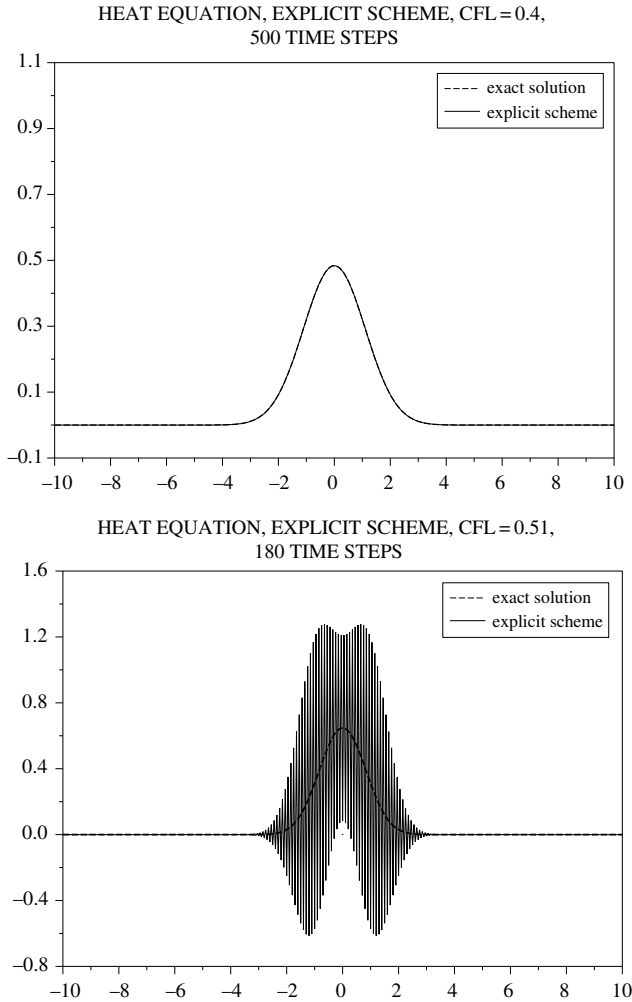


Figure 1.7. Explicit scheme with  $\nu\Delta t = 0.4(\Delta x)^2$  (top) and  $\nu\Delta t = 0.51(\Delta x)^2$  (bottom).

We shall briefly justify this stability condition (a more through analysis will be carried out in the next chapter). Rewriting the explicit scheme in the form

$$u_j^{n+1} = \frac{\nu\Delta t}{(\Delta x)^2} u_{j-1}^n + \left(1 - 2\frac{\nu\Delta t}{(\Delta x)^2}\right) u_j^n + \frac{\nu\Delta t}{(\Delta x)^2} u_{j+1}^n. \quad (1.32)$$

If the CFL condition is satisfied, then (1.32) shows that  $u_j^{n+1}$  is a convex combination of the values at the preceding time  $u_{j-1}^n, u_j^n, u_{j+1}^n$  (all of the coefficients on the right-hand side of (1.32) are positive and their sum is 1). In particular, if the initial data  $u_0$  is bounded by two constants  $m$  and  $M$  such that

$$m \leq u_j^0 \leq M \quad \text{for all } j \in \mathbb{Z},$$

then a recurrence easily shows that the same inequalities remain true for all time

$$m \leq u_j^n \leq M \quad \text{for all } j \in \mathbb{Z} \text{ and for all } n \geq 0. \quad (1.33)$$

Property (1.33) prevents the scheme from oscillating unboundedly: it is therefore stable subject to the CFL condition. Property (1.33) is called a discrete maximum principle: it is the **discrete** equivalent of the **continuous** maximum principle for exact solutions which we have seen in remark 1.2.10.

Suppose, on the other hand, the CFL condition is not satisfied, that is,

$$2\nu\Delta t > (\Delta x)^2,$$

then, for certain initial data the scheme is unstable (it may be stable for certain ‘exceptional’ initial data: for example, if  $u_0 \equiv 0$ !). Let us take the initial data defined by

$$u_j^0 = (-1)^j$$

which is uniformly bounded. A simple calculation shows that

$$u_j^n = (-1)^j \left( 1 - 4 \frac{\nu\Delta t}{(\Delta x)^2} \right)^n$$

which tends, in modulus, to infinity as  $n$  tends to infinity since  $1 - 4\nu\Delta t/(\Delta x)^2 < -1$ . The explicit scheme is therefore unstable if the CFL condition is not satisfied.

**Exercise 1.4.1** The aim of this exercise is to show that the implicit scheme (1.29), with  $V = 0$ , also satisfies the discrete maximum principle. We impose Dirichlet boundary conditions, that is, formula (1.29) is valid for  $1 \leq j \leq J$  and we fix  $u_0^n = u_{J+1}^n = 0$  for all  $n \in \mathbb{N}$ . Take two constants  $m \leq 0 \leq M$  such that  $m \leq u_j^0 \leq M$  for  $1 \leq j \leq J$ . Verify that we can uniquely calculate the  $u_j^{n+1}$  as a function of  $u_j^n$ . Show that for all time  $n \geq 0$  we again have the inequalities  $m \leq u_j^n \leq M$  for  $1 \leq j \leq J$  (without any condition on  $\Delta t$  and  $\Delta x$ ).

If we have illuminated the question of the stability of the explicit scheme a little, we have not said anything about its convergence, that is, its capacity to approximate the exact solution. We shall answer this question rigorously in the following chapter. We remark that stability is a necessary condition for convergence, but it is not sufficient. We shall be content for the moment with experimentally verifying the convergence of the scheme, that is, when the space and time steps become smaller and smaller,

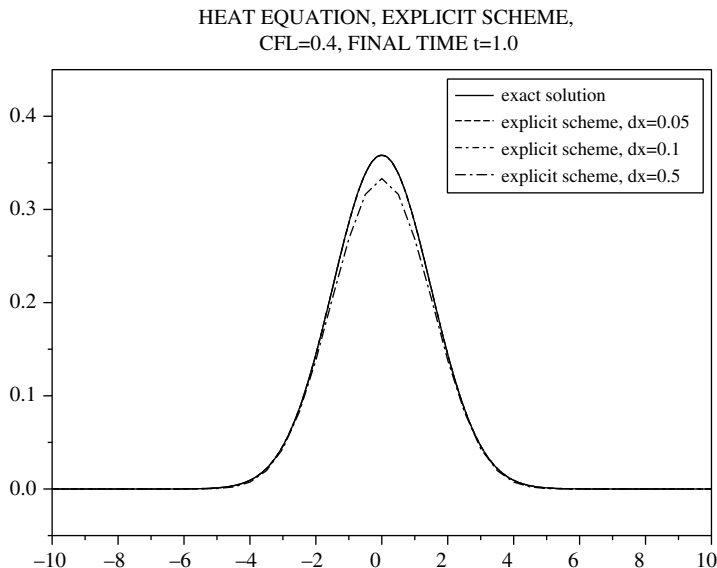


Figure 1.8. Explicit scheme with  $\nu\Delta t = 0.4(\Delta x)^2$  for various values of  $\Delta x$ .

the corresponding numerical solutions converge and their limit is the exact solution (we can check this as the exact solution is available). In Figure 1.8 we numerically verify that if we reduce the space step  $\Delta x$  (which has values 0.5, 0.1, and 0.05) and the time step  $\Delta t$  by keeping the ratio  $\nu\Delta t/(\Delta x)^2$  (the CFL number) constant, then the numerical solution becomes closer and closer to the exact solution. (The comparison is carried out at the same final time  $t = 1$ , therefore the number of time steps grows as the time step  $\Delta t$  decreases.) This process of ‘**numerical verification of convergence**’ is very simple and we should never hesitate to use it if nothing better is available (that is, if the theoretical convergence analysis is impossible or too difficult).

### 1.4.3 Numerical results for the advection equation

We shall carry out a second series of numerical experiments on the **convection–diffusion equation** (1.27) with a nonzero velocity  $V = 1$ . We take the same data as before and we choose the explicit scheme with  $\nu\Delta t = 0.4(\Delta x)^2$ . We look at the influence of the diffusion constant  $\nu$  (or the inverse of the Péclet number) on the stability of the scheme. Figure 1.9 shows that the scheme is stable when  $\nu = 1$ , unstable for  $\nu = 0.01$ , and that for the intermediate value  $\nu = 0.1$ , the scheme seems stable but the approximate solution is slightly different from the exact solution. Clearly, the smaller the inverse of the Péclet number  $\nu$  is, the more the convective term dominates the diffusive term. Consequently, the CFL condition (1.31), obtained when the velocity  $V$  is zero, is less and less valid as  $\nu$  decreases.

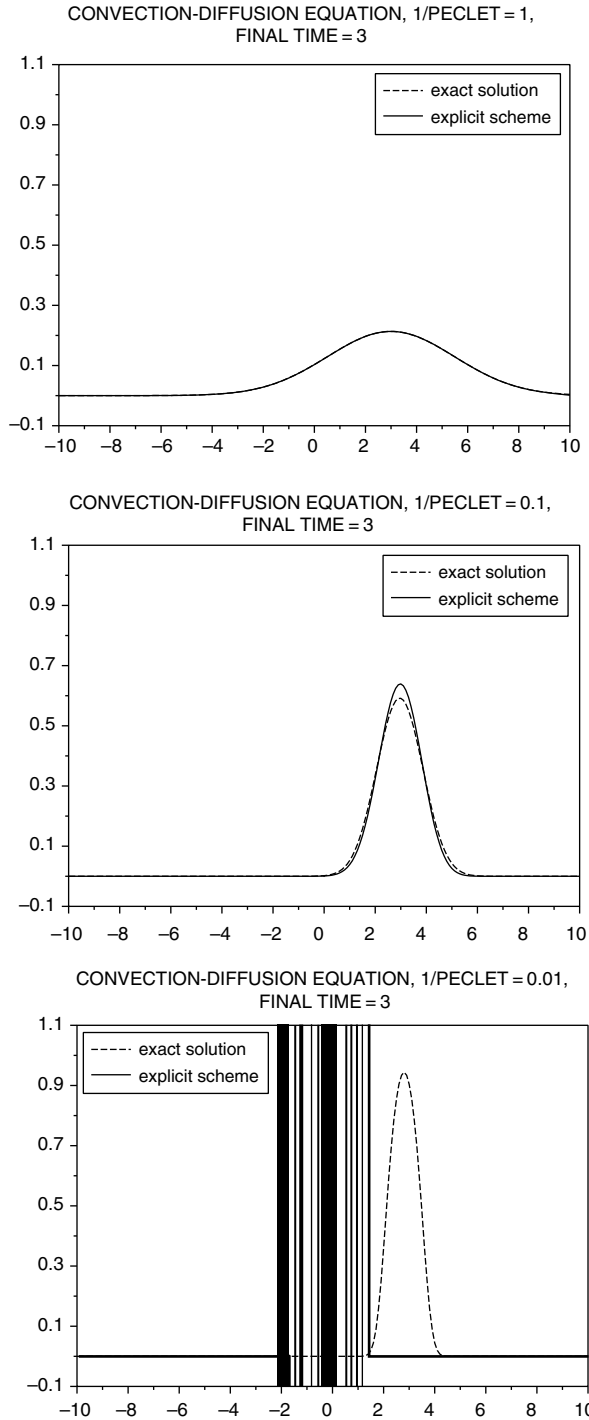


Figure 1.9. Explicit scheme for the convection–diffusion equation with  $\nu\Delta t = 0.4(\Delta x)^2$  and  $V = 1$ . At the top,  $\nu = 1$ , in the middle  $\nu = 0.1$ , and at the bottom  $\nu = 0.01$ .

To understand this phenomenon, we examine **the advection equation** which is obtained in the limit  $\nu = 0$ . We remark first that the CFL condition (1.31) is automatically satisfied when  $\nu = 0$  (whatever  $\Delta t$  and  $\Delta x$ ), which seems to contradict the experimental result at the bottom of Figure 1.9.

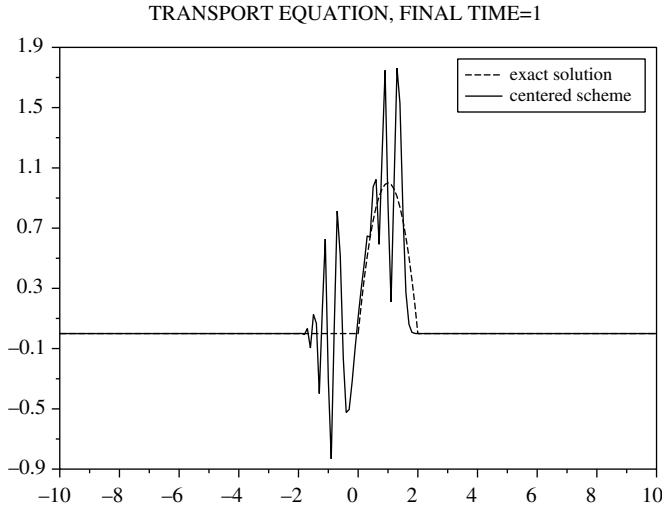


Figure 1.10. Explicit centred scheme for the advection equation with  $\Delta t = 0.9\Delta x$ ,  $V = 1$ ,  $\nu = 0$ .

For the advection equation (that is, (1.27) with  $\nu = 0$ ), the explicit scheme (1.30) may be rewritten

$$u_j^{n+1} = \frac{V\Delta t}{2\Delta x} u_{j-1}^n + u_j^n - \frac{V\Delta t}{2\Delta x} u_{j+1}^n. \quad (1.34)$$

This scheme leads to the oscillations in Figure 1.10 under the same experimental conditions as the bottom of Figure 1.9. We see that  $u_j^{n+1}$  is never (no matter what  $\Delta t$ ) a convex combination of  $u_{j-1}^n$ ,  $u_j^n$ , and  $u_{j+1}^n$ . Therefore, there cannot be a discrete maximum principle for this scheme, which is an additional indication of its instability (a rigorous proof will be given in lemma 2.3.1). This instability occurs because, in the explicit scheme (1.34), we have chosen to use a centred approximation to the convective term. We can, however, make this term one-sided as we have done for the time derivative. Two choices are possible: weighting to the right or left. The sign of the velocity  $V$  is crucial: here we assume that  $V > 0$  (a symmetric argument is possible if  $V < 0$ ). For  $V > 0$ , the weighting to the right is called **downwinding**: we obtain

$$V \frac{\partial u}{\partial x}(t_n, x_j) \approx V \frac{u_{j+1}^n - u_j^n}{\Delta x}$$

we try to find ‘information’ by following the current. This leads to a ‘disastrous’



downwind scheme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_j^n}{\Delta x} = 0 \quad (1.35)$$

which is as unstable as the centred scheme. On the other hand, the **upwinding** (which is to the left if  $V > 0$ ), looks for ‘information’ by going against the current

$$V \frac{\partial u}{\partial x}(t_n, x_j) \approx V \frac{u_j^n - u_{j-1}^n}{\Delta x}$$

leading to an explicit upwind scheme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \quad (1.36)$$

which gives the results of Figure 1.11. We verify easily that the scheme (1.36) is stable

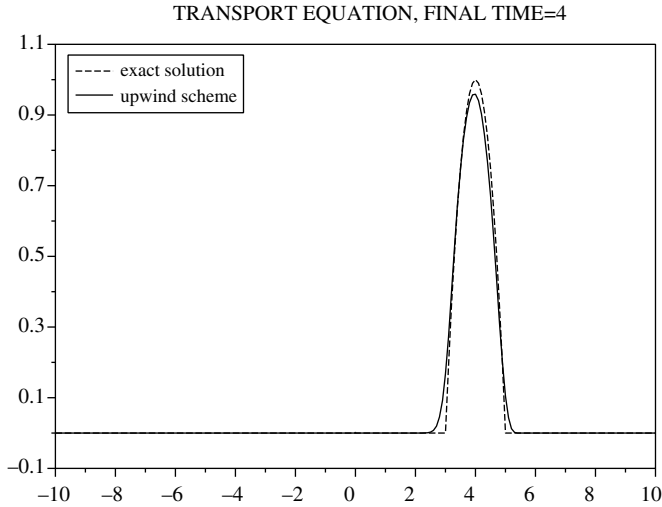


Figure 1.11. Explicit upwind scheme for the advection equation with  $\Delta t = 0.9$ ,  $\Delta x, V = 1$ .

under a new CFL condition (different from the preceding CFL condition (1.31))

$$|V|\Delta t \leq \Delta x. \quad (1.37)$$

Indeed, we can rewrite (1.36) in the form

$$u_j^{n+1} = \frac{V\Delta t}{\Delta x} u_{j-1}^n + \left(1 - \frac{V\Delta t}{\Delta x}\right) u_j^n,$$

which shows that, if condition (1.37) is satisfied,  $u_j^{n+1}$  is a convex combination of  $u_{j-1}^n$  and  $u_j^n$ . Consequently, the one-sided upwind scheme (1.36) satisfies a discrete maximum principle, which implies conditional stability. The idea of **upwinded methods is another major idea in numerical analysis**. It is particularly important in all fluid mechanics problems where it was first discovered, but it appears in many other models.

The conclusion of this study on the advection equation is that for the convection–diffusion model with a small diffusion constant  $\nu$ , we must upwind the convective term and obey the CFL condition (1.37) rather than (1.31). With this price we can improve the results of Figure 1.9.

**Exercise 1.4.2** Show that if (1.37) is not satisfied, the upwind scheme (1.36) for the advection equation is unstable for the initial data  $u_j^0 = (-1)^j$ .

**Exercise 1.4.3** Write an explicit scheme centred in space for the wave equation (1.18) in one space dimension and without source term. Specify how to start the iterations in time. Verify the existence of a discrete cone of dependence analogous to the continuous one shown in figure 1.3. Deduce that, if this scheme converges, the time and space steps must satisfy the (CFL-like) condition  $\Delta t \leq \Delta x$ .

The conclusions of this section are numerous and will feed the reflections of the subsequent chapter. First of all, all ‘reasonable’ numerical schemes do not work, far from it. We meet stability problems (without even considering convergence) which require us to analyse these schemes: this is the *raison d’être* of numerical analysis which reconciles practical objectives and theoretical studies. Finally, the ‘good’ numerical schemes must have a certain number of properties (for example, the discrete maximum principle, or upwinding) which are the expression (at the discrete level) of the physical properties or the mathematics of the PDE. **We cannot therefore skimp on a good understanding of the physical modelling and of the mathematical properties of the models if we want to have good numerical simulations.**

## 1.5 Remarks on mathematical models

We finish this chapter with a number of definitions which allow the reader to understand the terms in classical works on numerical analysis.

### 1.5.1 The idea of a well-posed problem

**Definition 1.5.1** We use the term **boundary value problem** to refer to a PDE equipped with boundary conditions on the entire boundary of the domain in which it is posed.

For example, the Laplacian (1.20) is a boundary value problem. Conversely, the ordinary differential equation

$$\begin{cases} \frac{dy}{dt} = f(t, y) & \text{for } 0 < t < T \\ y(t = 0) = y_0 \end{cases} \quad (1.38)$$

is not a boundary value problem as it is posed on an interval  $(0, T)$ , with  $0 < T \leq +\infty$ , it only has ‘boundary’ conditions at  $t = 0$  (and not at  $t = T$ ).

**Definition 1.5.2** *We say **Cauchy problem** to mean a PDE where, for at least one variable (usually time  $t$ ), the ‘boundary’ conditions are initial conditions (that is, only hold at the boundary  $t = 0$ , and not at  $t = T$ ).*

For example, the ordinary differential equation (1.38) is a Cauchy problem, but the Laplacian (1.20) is not (no matter which component of the space variable  $x$  we make to play the role of time).

Numerous models are, at the same time, boundary value problems and Cauchy problems. Thus, the heat flow equation (1.8) is a Cauchy problem with respect to the time variable  $t$  and a boundary value problem with respect to the space variable  $x$ . All the models we shall study in this course belong to one of these two categories of problem.

The fact that a mathematical model is a Cauchy problem or a boundary value problem does not automatically imply that it is a ‘good’ model. The expression **good model** is not used here in the sense of the physical relevance of the model and of its results, but in the sense of its mathematical coherence. As we shall see, this mathematical coherence is a necessary condition before we can consider numerical simulations and physical interpretations. The mathematician Jacques Hadamard gave a definition of what is a ‘good’ model, while speaking about **well-posed problems** (an ill-posed problem is the opposite of a well-posed problem). We denote by  $f$  the data (the right-hand side, the initial conditions, the domain, etc.),  $u$  the solution sought, and  $\mathcal{A}$  ‘the operator’ which acts on  $u$ . We are using abstract notation,  $\mathcal{A}$  denotes simultaneously the PDE and the type of initial or boundary conditions. The problem is therefore to find  $u$ , the solution of

$$\mathcal{A}(u) = f. \quad (1.39)$$

**Definition 1.5.3** *We say that problem (1.39) is **well-posed** if for all data  $f$  it has a unique solution  $u$ , and if this solution  $u$  depends continuously on the data  $f$ .*

Let us examine Hadamard’s definition in detail: it contains, in fact, three conditions for the problem to be well-posed. First, a solution must at least exist: this is the least we can ask of a model supposed to represent reality! Second, the solution must be unique: this is more delicate since, while it is clear that, if we want to predict tomorrow’s weather, it is better to have ‘sun’ or ‘rain’ (with an exclusive

‘or’) but not both with equal chance, there are other problems which ‘reasonably’ have several or an infinity of solutions. For example, problems involving finding the best route often have several solutions: to travel from the South to the North Pole then any meridian will do, likewise, to travel by plane from Paris to New York, your travel agency sometimes makes you go via Brussels or London, rather than directly, because it can be more economic. Hadamard excludes this type of problem from his definition since the multiplicity of solutions means that the model is indeterminate: to make the final choice between all of those that are best, we use another criterion (which has been ‘forgotten’ until now), for example, the most practical or most comfortable journey. This is a situation of current interest in applied mathematics: when a model has many solutions, we must add a selection criterion to obtain the ‘good’ solution (see, for a typical example, problems in gas dynamics [23]). Third, and this is the least obvious condition *a priori*, the solution must depend continuously on the data. At first sight, this seems a mathematical fantasy, but it is crucial from the perspective of **numerical approximation**. Indeed, numerically calculating an approximate solution of (1.39) amounts to perturbing the data (when continuous becomes discrete) and solving (1.39) for the perturbed data. If small perturbations of the data lead to large perturbations of the solution, there is no chance that the numerical approximation will be close to reality (or at least to the exact solution). Consequently, this continuous dependence of the solution on the data is an absolutely necessary condition for accurate numerical simulations. We note that this condition is also very important from the physical point of view since measuring apparatus will not give us absolute precision: if we are unable to distinguish between two close sets of data which can lead to very different phenomena, the model represented by (1.39) has no predictive value, and therefore is of almost no practical interest.

We finish by acknowledging that, at this level of generality, the definition (1.5.3) is a little fuzzy, and that to give it a precise mathematical sense we should say in which function spaces we put the data or look for the solution, and which norms or topologies we use for the continuity. It is not uncommon that changing the space (which can appear anodyne) implies very different properties of existence or uniqueness!

**Exercise 1.5.1** The point of this exercise is to show that the Cauchy problem for the Laplacian is ill-posed. Take a two-dimensional domain  $\Omega = (0, 1) \times (0, 2\pi)$ . We consider the following Cauchy problem in  $x$  and boundary value problem in  $y$

$$\begin{cases} -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0 & \text{in } \Omega \\ u(x, 0) = u(x, 2\pi) = 0 & \text{for } 0 < x < 1 \\ u(0, y) = 0, \frac{\partial u}{\partial x}(0, y) = -e^{-\sqrt{n}} \sin(ny) & \text{for } 0 < y < 2\pi \end{cases}$$

Verify that  $u(x, y) = (e^{-\sqrt{n}}/n) \sin(ny) \operatorname{sh}(nx)$  is a solution. Show that the initial condition and all its derivatives at  $x = 0$  converge uniformly to 0, while, for all  $x > 0$ , the solution  $u(x, y)$  and all its derivatives are unbounded as  $n$  tends to infinity.

### 1.5.2 Classification of PDEs

**Definition 1.5.4** *The order of a partial differential equation is the order of the highest derivative in the equation.*

For example, the Laplacian (1.20) is a second order equation, while the plate equation (1.24) is a fourth order equation. We often distinguish between the order with respect to the time variable  $t$  and with respect to the space variable  $x$ . Therefore, we say that heat flow equation (1.8) is first order in time and second order in space; likewise, the wave equation (1.18) is second order in space-time.

In order to understand the vocabulary often used with PDEs, that is, **elliptic**, **parabolic**, or **hyperbolic**, we shall briefly classify linear, second order PDEs acting on real functions of two real variables  $u(x, y)$  (we shall not carry out a systematic classification for all PDEs). Such an equation is written

$$a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial^2 u}{\partial x \partial y} + c \frac{\partial^2 u}{\partial y^2} + d \frac{\partial u}{\partial x} + e \frac{\partial u}{\partial y} + fu = g. \quad (1.40)$$

For simplicity we assume that the coefficients  $a, b, c, d, e, f$  are constant.

**Definition 1.5.5** *We say that the equation (1.40) is elliptic if  $b^2 - 4ac < 0$ , parabolic if  $b^2 - 4ac = 0$ , and hyperbolic if  $b^2 - 4ac > 0$ .*

The origin of this vocabulary is in the classification of conic sections, from which Definition 1.5.5 is copied. Indeed, it is well-known that the second degree equation

$$ax^2 + bxy + cy^2 + dx + ey + f = 0$$

defines a plane curve which is (except in some degenerate cases) an ellipse if  $b^2 - 4ac < 0$ , a parabola if  $b^2 - 4ac = 0$ , and a hyperbola if  $b^2 - 4ac > 0$ .

If we apply Definition 1.5.5 to the various second order models we have stated in this chapter (replacing the variables  $(x, y)$  by the variables  $(t, x)$  in one space dimension), we conclude that **the heat flow equation is parabolic** (as is the convection–diffusion equation), that **the Laplacian is elliptic**, and that **the wave equation is hyperbolic**. A suitable generalisation of this definition allows us to check that the advection equation is hyperbolic, and that the Stokes, elasticity, and plate equations are elliptic. In general, stationary problems (independent of time) are modelled by elliptic PDEs, while evolution problems are modelled by parabolic or hyperbolic PDEs.

We shall see later that boundary value problems are well posed for elliptic PDEs, while problems which are Cauchy in time and boundary value problems in space are well-posed for parabolic or hyperbolic PDEs. There are therefore important differences in behaviour between these two types of equation.

**Remark 1.5.6** The elliptic, hyperbolic or parabolic character of the equations (1.40) is not modified by a change of variable. Let  $(x, y) \rightarrow (X, Y)$  be such a change of variable which is nonsingular, that is, its Jacobian  $J = X_x Y_y - X_y Y_x$  is not zero (denoting by  $Z_z$  the derivative of  $Z$  with respect to  $z$ ). A simple but tedious calculation shows that (1.40) becomes

$$A \frac{\partial^2 u}{\partial X^2} + B \frac{\partial^2 u}{\partial X \partial Y} + C \frac{\partial^2 u}{\partial Y^2} + D \frac{\partial u}{\partial X} + E \frac{\partial u}{\partial Y} + Fu = G,$$

with  $A = aX_x^2 + bX_xX_y + cX_y^2$ ,  $B = 2aX_xY_x + b(X_xY_y + X_yY_x) + 2cX_yY_y$ ,  $C = aY_x^2 + bY_xY_y + cY_y^2$ , and we verify that  $B^2 - 4AC = J^2(b^2 - 4ac)$ . In particular, a suitable change of variables allows us to simplify the PDE (1.40) and return it to its ‘canonical’ form. Thus, any elliptic equation can be reduced to the Laplacian  $\frac{\partial^2}{\partial X^2} + \frac{\partial^2}{\partial Y^2}$ , any parabolic equation to the heat flow equation  $\frac{\partial}{\partial X} - \frac{\partial^2}{\partial Y^2}$ , and any hyperbolic equation to the wave equation  $\frac{\partial^2}{\partial X^2} - \frac{\partial^2}{\partial Y^2}$ . •

**Remark 1.5.7** It is well-known that the general conic equation has a number of degenerate cases when it no longer describes a cone but a set of lines. The same situation can hold with the PDE (1.40). For example, the equation  $\frac{\partial^2 u}{\partial x^2} = 1$  with  $a = 1$  and  $b = c = d = e = f = 0$  is not parabolic in two dimensions (even though  $b^2 - 4ac = 0$ ) but elliptic in one dimension (the variable  $y$  plays no role here). It is therefore necessary to be careful before deciding on the type of these ‘degenerate’ equations. •

*This page intentionally left blank*

# 2 Finite difference method

---

## 2.1 Introduction

In this chapter we analyse numerical schemes of finite differences. We define the **stability** and **consistency** of a scheme and show that, for linear, constant coefficient, partial differential equations, stability plus consistency of a scheme implies its **convergence**.

The plan of the chapter is the following. Section 2.2 treats the case of the heat equation introduced in Chapter 1. Section 2.3 generalizes the preceding results to the case of the wave equation or the advection equation. One of the aims of this chapter is the construction and analysis of finite differences schemes for much more general models. The reader should not be afraid of extending the concepts presented here to his preferred model and to construct original numerical schemes.

We finish this introduction by saying that the finite difference method is one of the oldest methods of numerical approximation which is still used in applications, such as wave propagation (seismic or electromagnetic) or compressible fluid mechanics. For other applications, such as solid mechanics or incompressible fluids, we often prefer the finite element method. Nevertheless, many concepts in finite differences are found in other numerical methods. Thus, the numerical schemes of Chapter 8 will combine finite elements for the space discretization and finite differences for the time discretization. The generality and simplicity of the finite difference method motivates our detailed study at the beginning of this work.



## 2.2 Finite differences for the heat equation

### 2.2.1 Various examples of schemes

We restrict ourselves to one space dimension and we refer to Section 2.2.6 for the case of several space dimensions. We consider the heat equation in the bounded domain  $(0, 1)$

$$\begin{cases} \frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} = 0 & \text{for } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ u(0, x) = u_0(x) & \text{for } x \in (0, 1). \end{cases} \quad (2.1)$$

To discretize the domain  $(0, 1) \times \mathbb{R}^+$ , we introduce a space step  $\Delta x = 1/(N+1) > 0$  (with  $N$  a positive integer) and a time step  $\Delta t > 0$ , and we define the nodes of a regular mesh

$$(t_n, x_j) = (n\Delta t, j\Delta x) \quad \text{for } n \geq 0, j \in \{0, 1, \dots, N+1\}.$$

We denote by  $u_j^n$  the value of a discrete approximate solution at the point  $(t_n, x_j)$ , and  $u(t, x)$  the exact solution of (2.1). The initial data is discretized by

$$u_j^0 = u_0(x_j) \quad \text{for } j \in \{0, 1, \dots, N+1\}.$$

The boundary conditions of (2.1) can be of several types, but their choice is not involved in the definition of the schemes. Here, we use Dirichlet boundary conditions

$$u(t, 0) = u(t, 1) = 0 \quad \text{for all } t \in \mathbb{R}_*^+$$

which imply

$$u_0^n = u_{N+1}^n = 0 \quad \text{for all } n > 0.$$

Consequently, at each time step we have to calculate the values  $(u_j^n)_{1 \leq j \leq N}$  which form a vector of  $\mathbb{R}^N$ . We now give several possible schemes for the heat equation (2.1). All of them are defined by  $N$  equations (at each point  $x_j$ ,  $1 \leq j \leq N$ ) which allow us to calculate the  $N$  values  $u_j^n$ . In Chapter 1 we have already talked of the **explicit scheme**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0 \quad (2.2)$$

for  $n \geq 0$  and  $j \in \{1, \dots, N\}$ , and also of the **implicit scheme**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \nu \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} = 0. \quad (2.3)$$

It is easy to verify that the implicit scheme (2.3) is well defined, that is, we can calculate the values  $u_j^{n+1}$  as a function of the  $u_j^n$ : in effect, we must invert the square

tridiagonal matrix of dimension  $N$

$$\begin{pmatrix} 1+2c & -c & & & 0 \\ -c & 1+2c & -c & & \\ & \ddots & \ddots & \ddots & \\ & & -c & 1+2c & -c \\ 0 & & & -c & 1+2c \end{pmatrix} \quad \text{with } c = \frac{\nu\Delta t}{(\Delta x)^2}, \quad (2.4)$$

which is easily verified to be positive definite, therefore invertible. By making a convex combination of (2.2) and (2.3), for  $0 \leq \theta \leq 1$ , we obtain the  **$\theta$ -scheme**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \theta\nu \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} + (1-\theta)\nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \quad (2.5)$$

We recover the explicit scheme (2.2) if  $\theta = 0$ , and the implicit scheme (2.3) if  $\theta = 1$ . The  $\theta$ -scheme (2.5) is implicit when  $\theta \neq 0$ . For the value  $\theta = 1/2$ , we obtain the **Crank–Nicolson scheme**. Another implicit scheme, called the six point scheme, is given by

$$\begin{aligned} & \frac{u_{j+1}^{n+1} - u_{j+1}^n}{12\Delta t} + \frac{5(u_j^{n+1} - u_j^n)}{6\Delta t} + \frac{u_{j-1}^{n+1} - u_{j-1}^n}{12\Delta t} \\ & + \nu \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{2(\Delta x)^2} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{2(\Delta x)^2} = 0. \end{aligned} \quad (2.6)$$

**Exercise 2.2.1** Show that the scheme (2.6) is nothing more than the  $\theta$ -scheme with  $\theta = 1/2 - (\Delta x)^2/12\nu\Delta t$ .

All the schemes above are called **two level** since they only involve two time indices. We can construct multilevel schemes: the most popular have three levels. In addition to the (unstable) Richardson scheme seen in Chapter 1, we cite the **DuFort–Frankel scheme**

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + \nu \frac{-u_{j-1}^n + u_j^{n+1} + u_j^{n-1} - u_{j+1}^n}{(\Delta x)^2} = 0, \quad (2.7)$$

the **Gear scheme**

$$\frac{3u_j^{n+1} - 4u_j^n + u_j^{n-1}}{2\Delta t} + \nu \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} = 0. \quad (2.8)$$

We have too many schemes! And the list above is not exhaustive! One of the aims of numerical analysis is to compare and to choose the best schemes following criteria of accuracy, cost, or robustness.

**Remark 2.2.1** If there is a right-hand side  $f(t, x)$  in the heat equation (2.1), then the schemes are modified by replacing zero in the right-hand side by a consistent approximation of  $f(t, x)$  at the point  $(t_n, x_j)$ . For example, if we choose the approximation  $f(t_n, x_j)$ , the explicit scheme (2.2) becomes

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = f(t_n, x_j).$$

•

**Remark 2.2.2** The schemes above are written compactly, that is, they involve a finite number of values  $u_j^n$ . The set of the couples  $(n', j')$  which appear in the discrete equation at the point  $(n, j)$  is called the **stencil** of the scheme. In general, the larger the stencil, the more costly and difficult it is to program the scheme (partly because of the ‘boundary effects’, that is, the case where some of the couples  $(n', j')$  leave the domain of calculation).

•

**Remark 2.2.3** We can replace the Dirichlet boundary conditions in (2.1) by Neumann boundary conditions, or by periodic (or other) boundary conditions. We start by describing two different ways of discretizing Neumann conditions

$$\frac{\partial u}{\partial x}(t, 0) = 0 \quad \text{and} \quad \frac{\partial u}{\partial x}(t, 1) = 0.$$

First, we can write

$$\frac{u_1^n - u_0^n}{\Delta x} = 0 \quad \text{and} \quad \frac{u_{N+1}^n - u_N^n}{\Delta x} = 0$$

which allow us to eliminate the values  $u_0^n$  and  $u_{N+1}^n$  and only to calculate the  $N$  values  $(u_j^n)_{1 \leq j \leq N}$ . This discretization of the Neumann condition is only first order. If the scheme is second order, this causes a loss of accuracy close to the boundary. This is why we propose another discretization (of second order)

$$\frac{u_1^n - u_{-1}^n}{2\Delta x} = 0 \quad \text{and} \quad \frac{u_{N+2}^n - u_N^n}{2\Delta x} = 0$$

which is more accurate, but needs us to add 2 ‘fictitious points’  $x_{-1}$  and  $x_{N+2}$ . We eliminate the values  $u_{-1}^n$  and  $u_{N+2}^n$ , corresponding to these fictitious points, and there now remains  $N + 2$  values to calculate, that is,  $(u_j^n)_{0 \leq j \leq N+1}$ .

On the other hand, periodic boundary conditions are written

$$u(t, x + 1) = u(t, x) \quad \text{for all } x \in [0, 1], \quad t \geq 0.$$

These are discretized by the equations  $u_0^n = u_{N+1}^n$  for all  $n \geq 0$ , and more generally  $u_j^n = u_{N+1+j}^n$ .

•

## 2.2.2 Consistency and accuracy

Of course, the formulas of the schemes above are not chosen by accident: they follow from an approximation of the equation by Taylor expansion as we have explained in Chapter 1. To formalize this approximation of the partial differential equation by finite differences, we introduce the ideas of **consistency** and of **accuracy**. Although for the moment we only consider the heat equation (2.1), we shall give a definition of the consistency which is valid for every partial differential equation which we write  $F(u) = 0$ . We remark that  $F(u)$  is notation for a function of  $u$  and its partial derivatives at every point  $(t, x)$ . Generally a finite difference scheme is defined, for all possible indices  $n, j$ , by the formula

$$F_{\Delta t, \Delta x} \left( \{u_{j+k}^{n+m}\}_{m^- \leq m \leq m^+, k^- \leq k \leq k^+} \right) = 0 \quad (2.9)$$

where the integers  $m^-, m^+, k^-, k^+$  define the width of the stencil of the scheme (see remark 2.2.2).

**Definition 2.2.4** *The finite difference scheme (2.9) is called consistent with the partial differential equation  $F(u) = 0$ , if, for every sufficiently regular solution  $u(t, x)$  of this equation, the truncation error of the scheme, defined by*

$$F_{\Delta t, \Delta x} \left( \{u(t + m\Delta t, x + k\Delta x)\}_{m^- \leq m \leq m^+, k^- \leq k \leq k^+} \right), \quad (2.10)$$

*tends to zero, uniformly with respect to  $(t, x)$ , as  $\Delta t$  and  $\Delta x$  tend to zero independently.*

*Further, we say that the scheme has accuracy of order  $p$  in space and order  $q$  in time if the truncation error (2.10) tends to zero as  $\mathcal{O}((\Delta x)^p + (\Delta t)^q)$  when  $\Delta t$  and  $\Delta x$  tend to zero.*

**Remark 2.2.5** We must take care with the formula (2.9) since there is a small ambiguity in the definition of the scheme. Indeed, we can always multiply any formula by a sufficiently high power of  $\Delta t$  and  $\Delta x$  so that the truncation error tends to zero. This will make any scheme consistent! To avoid this problem, we always assume that the formula  $F_{\Delta t, \Delta x}(\{u_{j+k}^{n+m}\}) = 0$  has been written so that, for a regular function  $u(t, x)$  which is not a solution of the equation  $F(u) = 0$ , the limit of the truncation error is not zero. •

Concretely we calculate the truncation error of a scheme by replacing  $u_{j+k}^{n+m}$  in formula (2.9) by  $u(t + m\Delta t, x + k\Delta x)$ . As an application of the definition 2.2.4, we shall show the following lemma.

**Lemma 2.2.6** *The explicit scheme (2.2) is consistent, accurate with order 1 in time and 2 in space. Further, if we choose to keep the ratio  $\nu\Delta t/(\Delta x)^2 = 1/6$  constant, then this scheme is accurate with order 2 in time and 4 in space.*

**Remark 2.2.7** In the second phrase of the statement of lemma 2.2.6 we slightly modified the definition of the consistency by specifying the ratio of  $\Delta t$  and  $\Delta x$  as they tend to zero. This allows us to take advantage of potential cancellation between terms in the truncation error. In practice, we see such improvements in the accuracy if we adopt a good relationship between the terms  $\Delta t$  and  $\Delta x$ . •

**Proof.** Let  $v(t, x)$  be a function of class  $\mathcal{C}^6$ . By Taylor expansion around the point  $(t, x)$ , we calculate the truncation error of the scheme (2.2)

$$\begin{aligned} & \frac{v(t + \Delta t, x) - v(t, x)}{\Delta t} + \nu \frac{-v(t, x - \Delta x) + 2v(t, x) - v(t, x + \Delta x)}{(\Delta x)^2} \\ &= (v_t - \nu v_{xx}) + \frac{\Delta t}{2} v_{tt} - \frac{\nu(\Delta x)^2}{12} v_{xxxx} + \mathcal{O}((\Delta t)^2 + (\Delta x)^4), \end{aligned}$$

where  $v_t, v_x$  denote the partial derivatives of  $v$ . If  $v$  is a solution of the heat equation (2.1), we thus easily obtain the consistency as well as accuracy of order 1 in time and 2 in space. If further we assume that  $\nu \Delta t / (\Delta x)^2 = 1/6$ , then the terms in  $\Delta t$  and  $(\Delta x)^2$  cancel since  $v_{tt} = \nu v_{txx} = \nu^2 v_{xxxx}$ .  $\square$

Scheme	Truncation error	Stability
Explicit (2.2)	$\mathcal{O}(\Delta t + (\Delta x)^2)$	Stable in $L^2$ and $L^\infty$ for CFL condition $2\nu\Delta t \leq (\Delta x)^2$
Implicit (2.3)	$\mathcal{O}(\Delta t + (\Delta x)^2)$	Stable in $L^2$ and $L^\infty$
Crank–Nicolson (2.5) (with $\theta = 1/2$ )	$\mathcal{O}((\Delta t)^2 + (\Delta x)^2)$	Stable in $L^2$
$\theta$ -scheme (2.5) (with $\theta \neq 1/2$ )	$\mathcal{O}(\Delta t + (\Delta x)^2)$	Stable in $L^2$ for CFL condition $2(1 - 2\theta)\nu\Delta t \leq (\Delta x)^2$
Six point scheme (2.6)	$\mathcal{O}((\Delta t)^2 + (\Delta x)^4)$	Stable in $L^2$
DuFort–Frankel (2.7)	$\mathcal{O}((\Delta t/\Delta x)^2 + (\Delta x)^2)$	Stable in $L^2$ for CFL condition $\Delta t/(\Delta x)^2$ bounded
Gear (2.8)	$\mathcal{O}((\Delta t)^2 + (\Delta x)^2)$	Stable in $L^2$

Table 2.1. Truncation errors and stability of various schemes for the heat equation

**Exercise 2.2.2** For each of the schemes of Section 2.2.1, verify that the truncation error is of the type stated in Table 2.1. (We remark that all these schemes are consistent except for DuFort–Frankel.)

### 2.2.3 Stability and Fourier analysis

In Chapter 1 we introduced the stability of finite differences schemes without giving a precise definition. We have explained that, numerically, instability is shown by unbounded oscillations of the numerical solution. It is therefore time to give a mathematical definition of stability. For this we need to define a norm for the numerical solution  $u^n = (u_j^n)_{1 \leq j \leq N}$ . We take the classical norms on  $\mathbb{R}^N$  which we scale by the space step  $\Delta x$ :

$$\|u^n\|_p = \left( \sum_{j=1}^N \Delta x |u_j^n|^p \right)^{1/p} \quad \text{for } 1 \leq p \leq +\infty, \quad (2.11)$$

where the limiting case  $p = +\infty$  should be understood in the sense  $\|u^n\|_\infty = \max_{1 \leq j \leq N} |u_j^n|$ . We remark that the norm defined therefore depends on  $\Delta x$  through the weighting but also on the integer  $N$  since  $\Delta x = 1/(N+1)$ . Thanks to the weighting by  $\Delta x$ , the norm  $\|u^n\|_p$  is identical to the norm  $L^p(0,1)$  for piecewise constant functions over the subintervals  $[x_j, x_{j+1}[$  of  $[0,1]$ . Often, we shall call this the ‘ $L^p$  norm’. In practice, we most often use the norms corresponding to the values  $p = 2, +\infty$ .

**Definition 2.2.8** *A finite difference scheme is called **stable** for the norm  $\|\cdot\|$ , defined by (2.11), if there exists a constant  $K > 0$  independent of  $\Delta t$  and  $\Delta x$  (as these values tend to zero) such that*

$$\|u^n\| \leq K \|u^0\| \quad \text{for all } n \geq 0, \quad (2.12)$$

for arbitrary initial data  $u^0$ .

If (2.12) only hold for steps  $\Delta t$  and  $\Delta x$  defined by certain inequalities, we say that the scheme is **conditionally stable**.

**Remark 2.2.9** Since all norms are equivalent in  $\mathbb{R}^N$ , the hasty reader might believe that stability with respect to one norm implies stability with respect to all norms. Unfortunately, this is not true and there exist schemes which are stable with respect to one norm but not with respect to another (see later the example of the Lax–Wendroff scheme in exercises 2.3.2 and 2.3.3). In effect, the crucial point in the definition 2.2.8 is that the bound is uniform with respect to  $\Delta x$  while the norms (2.11) depend on  $\Delta x$ . •

**Definition 2.2.10** *A finite difference scheme is called **linear** if its formula  $F_{\Delta t, \Delta x}(\{u_{j+k}^{n+m}\}) = 0$  is linear with respect to its arguments  $u_{j+k}^{n+m}$ .*

The stability of a two level linear scheme is very easy to interpret. Indeed, by linearity every two level linear scheme can be written in the condensed form

$$u^{n+1} = Au^n, \quad (2.13)$$

where  $A$  is a linear operator (a matrix, called the iteration matrix) from  $\mathbb{R}^N$  into  $\mathbb{R}^N$ . For example, for the explicit scheme (2.2) the matrix  $A$  becomes

$$\begin{pmatrix} 1-2c & c & & & 0 \\ c & 1-2c & c & & \\ & \ddots & \ddots & \ddots & \\ & & c & 1-2c & c \\ 0 & & & c & 1-2c \end{pmatrix} \quad \text{with } c = \frac{\nu \Delta t}{(\Delta x)^2}, \quad (2.14)$$

while for the implicit scheme (2.3) the matrix  $A$  is the inverse of the matrix (2.4). With the help of this iteration matrix, we have  $u^n = A^n u^0$  (take care, the notation  $A^n$  denotes the  $n$ th power of  $A$ ), and consequently the stability of the scheme is equivalent to

$$\|A^n u^0\| \leq K \|u^0\| \quad \forall n \geq 0, \quad \forall u^0 \in \mathbb{R}^N.$$

Introducing the subordinate matrix norm (see definition 13.1.1)

$$\|M\| = \sup_{u \in \mathbb{R}^N, u \neq 0} \frac{\|Mu\|}{\|u\|},$$

the stability of the scheme is equivalent to

$$\|A^n\| \leq K \quad \forall n \geq 0, \quad (2.15)$$

which is the same as saying the sequence of the powers of  $A$  is bounded.

### Stability in the $L^\infty$ norm

The stability in the  $L^\infty$  norm is closely linked with the discrete maximum principle which we have seen in Chapter 1. Let us recall the definition of this principle.

**Definition 2.2.11** *A finite difference scheme satisfies the **discrete maximum principle** if for all  $n \geq 0$  and all  $1 \leq j \leq N$  we have*

$$\min \left( 0, \min_{0 \leq j \leq N+1} u_j^0 \right) \leq u_j^n \leq \max \left( 0, \max_{0 \leq j \leq N+1} u_j^0 \right)$$

for arbitrary initial data  $u^0$ .

**Remark 2.2.12** In definition 2.2.11 the inequalities take account not only of the minimum and maximum of  $u^0$  but also of zero which is the value imposed on the boundary by the Dirichlet boundary conditions. This is necessary if the initial data  $u^0$  does not satisfy the Dirichlet boundary conditions (which is not required), and superfluous in the complementary case. •

As we have seen in Chapter 1 (see (1.33) and exercise 1.4.1), the discrete maximum principle allows us to prove the following lemma.

**Lemma 2.2.13** *The explicit scheme (2.2) is stable in the  $L^\infty$  norm if and only if the CFL condition  $2\nu\Delta t \leq (\Delta x)^2$  is satisfied. The implicit scheme (2.3) is stable in the  $L^\infty$  norm no matter what the time step  $\Delta t$  and space step  $\Delta x$  (we say that it is unconditionally stable).*

**Exercise 2.2.3** Show that the Crank–Nicolson scheme (2.5) (with  $\theta = 1/2$ ) is stable in the  $L^\infty$  norm if  $\nu\Delta t \leq (\Delta x)^2$ , and that the DuFort–Frankel scheme (2.7) is stable in the  $L^\infty$  norm if  $2\nu\Delta t \leq (\Delta x)^2$ .

### Stability in the $L^2$ norm

Many schemes do not satisfy the discrete maximum principle but are nevertheless ‘good’ schemes. For this, we must verify the stability in a norm other than the  $L^\infty$  norm. The  $L^2$  norm lends itself very well to the study of stability thanks to the very powerful tool of Fourier analysis which we now present. To do this, we assume from now on that the boundary conditions for the heat equation are **periodic boundary conditions**, which are written  $u(t, x+1) = u(t, x)$  for all  $x \in [0, 1]$  and all  $t \geq 0$ . For numerical schemes, these lead to the equations  $u_0^n = u_{N+1}^n$  for all  $n \geq 0$ , and more generally  $u_j^n = u_{N+1+j}^n$ . We therefore have to calculate  $N+1$  values  $u_j^n$ .

With each vector  $u^n = (u_j^n)_{0 \leq j \leq N}$  we associate a function  $u^n(x)$ , piecewise constant, periodic with period 1, defined on  $[0, 1]$  by

$$u^n(x) = u_j^n \quad \text{if } x_{j-1/2} < x < x_{j+1/2}$$

with  $x_{j+1/2} = (j+1/2)\Delta x$  for  $0 \leq j \leq N$ ,  $x_{-1/2} = 0$ , and  $x_{N+1+1/2} = 1$ . The function  $u^n(x)$  belongs to  $L^2(0, 1)$ . Now, from Fourier analysis, every function of  $L^2(0, 1)$  can be decomposed into a Fourier sum (see [4], [35], [38]). More precisely we have

$$u^n(x) = \sum_{k \in \mathbb{Z}} \hat{u}^n(k) \exp(2i\pi kx), \quad (2.16)$$

with  $\hat{u}^n(k) = \int_0^1 u^n(x) \exp(-2i\pi kx) dx$  and the Plancherel formula

$$\int_0^1 |u^n(x)|^2 dx = \sum_{k \in \mathbb{Z}} |\hat{u}^n(k)|^2. \quad (2.17)$$

We remark that even if  $u^n$  is a real function, the coefficients  $\hat{u}^n(k)$  of the Fourier series are complex. An important property for the Fourier transform of periodic functions is the following: if we denote by  $v^n(x) = u^n(x + \Delta x)$ , then  $\hat{v}^n(k) = \hat{u}^n(k) \exp(2i\pi k\Delta x)$ .

Let us now explain the method using the example of the explicit scheme (2.2). Under our notation, we can rewrite this scheme, for  $0 \leq x \leq 1$ ,

$$\frac{u^{n+1}(x) - u^n(x)}{\Delta t} + \nu \frac{-u^n(x - \Delta x) + 2u^n(x) - u^n(x + \Delta x)}{(\Delta x)^2} = 0.$$



By application of the Fourier transform, this becomes

$$\hat{u}^{n+1}(k) = \left(1 - \frac{\nu\Delta t}{(\Delta x)^2} (-\exp(-2i\pi k\Delta x) + 2 - \exp(2i\pi k\Delta x))\right) \hat{u}^n(k).$$

In other words,

$$\hat{u}^{n+1}(k) = A(k)\hat{u}^n(k) = A(k)^{n+1}\hat{u}^0(k) \quad \text{with } A(k) = 1 - \frac{4\nu\Delta t}{(\Delta x)^2}(\sin(\pi k\Delta x))^2.$$

For  $k \in \mathbb{Z}$ , the Fourier coefficient  $\hat{u}^n(k)$  is bounded as  $n$  tends to infinity if and only if the amplification factor satisfies  $|A(k)| \leq 1$ , that is,

$$2\nu\Delta t(\sin(\pi k\Delta x))^2 \leq (\Delta x)^2. \quad (2.18)$$

If the CFL condition (1.31), that is,  $2\nu\Delta t \leq (\Delta x)^2$ , is satisfied, then inequality (2.18) is true for every Fourier mode  $k \in \mathbb{Z}$ , and by the Plancherel formula we deduce

$$\|u^n\|_2^2 = \int_0^1 |u^n(x)|^2 dx = \sum_{k \in \mathbb{Z}} |\hat{u}^n(k)|^2 \leq \sum_{k \in \mathbb{Z}} |\hat{u}^0(k)|^2 = \int_0^1 |u^0(x)|^2 dx = \|u^0\|_2^2,$$

which is nothing other than the  $L^2$  stability of the explicit scheme. If the CFL condition is not satisfied, the scheme is unstable. In effect, it is enough to choose  $\Delta x$  (possibly sufficiently small) and  $k_0$  (sufficiently large) and initial data with only one nonzero Fourier component  $\hat{u}^0(k_0) \neq 0$  with  $\pi k_0 \Delta x \approx \pi/2$  (modulo  $\pi$ ) in such a way that  $|A(k_0)| > 1$ . We have therefore proved the following lemma.

**Lemma 2.2.14** *The explicit scheme (2.2) is stable in the  $L^2$  norm if and only if the CFL condition  $2\nu\Delta t \leq (\Delta x)^2$  is satisfied.*

In the same way we shall prove the stability of the implicit scheme.

**Lemma 2.2.15** *The implicit scheme (2.3) is stable in the  $L^2$  norm.*

**Remark 2.2.16** For explicit (2.2) and implicit (2.3) schemes the  $L^2$  stability condition is the same as that of the  $L^\infty$  stability. This is not always the case for other schemes. •

**Proof.** Similar reasoning to that used for the explicit scheme leads, for  $0 \leq x \leq 1$ , to

$$\frac{u^{n+1}(x) - u^n(x)}{\Delta t} + \nu \frac{-u^{n+1}(x - \Delta x) + 2u^{n+1}(x) - u^{n+1}(x + \Delta x)}{(\Delta x)^2} = 0,$$

and by application of the Fourier transform

$$\hat{u}^{n+1}(k) \left(1 + \frac{\nu\Delta t}{(\Delta x)^2} (-\exp(-2i\pi k\Delta x) + 2 - \exp(2i\pi k\Delta x))\right) = \hat{u}^n(k).$$

In other words,

$$\hat{u}^{n+1}(k) = A(k)\hat{u}^n(k) = A(k)^{n+1}\hat{u}^0(k) \text{ with } A(k) = \left(1 + \frac{4\nu\Delta t}{(\Delta x)^2}(\sin(\pi k\Delta x))^2\right)^{-1}.$$

As  $|A(k)| \leq 1$  for all Fourier modes  $k$ , the Plancherel formula gives us the  $L^2$  stability of the scheme.  $\square$

**Remark 2.2.17** The Fourier analysis relies on the choice of periodic boundary conditions. We can also carry it out if the partial differential equation holds over all  $\mathbb{R}$  instead of  $[0, 1]$  (we have then to deal with a Fourier integral instead of a Fourier series). Nevertheless, it is not very realistic to talk about a numerical scheme over all  $\mathbb{R}$  since this implies an infinite number of values  $u_j^n$  at each time step  $n$  when a computer can only treat a finite number of values.

The  $L^2$  stability can also be proved in the case of Dirichlet boundary conditions. We must then adapt the ideas of Fourier analysis. For example, what replaces the Fourier transform in this case is the decomposition over a basis of eigenvectors of the iteration matrix (2.13) which allows us to move from the vector  $u^n$  to the vector  $u^{n+1}$ .  $\bullet$

**Remark 2.2.18 (Essential from a practical point of view)** Let us give a ‘recipe’ for Fourier analysis to prove the  $L^2$  stability of a scheme. We put Fourier modes into the scheme

$$u_j^n = A(k)^n \exp(2i\pi k x_j) \quad \text{with } x_j = j\Delta x,$$

and we deduce the value of the amplification factor  $A(k)$ . Recall that, for the moment, we restrict ourselves to the scalar case, that is,  $A(k)$  is a complex number in  $\mathbb{C}$ . The inequality

$$|A(k)| \leq 1 \quad \text{for all modes } k \in \mathbb{Z} \tag{2.19}$$

is called the **Von Neumann stability condition**. If the Von Neumann stability condition is satisfied (with possibly restrictions on  $\Delta t$  and  $\Delta x$ ), then the scheme is stable for the  $L^2$  norm, if not it is unstable.

In general, a stable (and consistent) scheme is convergent (see Section 2.2.4). In practice, an unstable scheme is totally ‘useless’. In effect, even if we start from initial data specially designed so that none of the unstable Fourier modes are excited, the inevitable rounding errors will create nonzero components (although very small) of the solution in the unstable modes. The exponential increase of the unstable modes implies that after only a few time steps these ‘small’ modes become ‘enormous’ and completely pollute the rest of the numerical solution.  $\bullet$

**Exercise 2.2.4** Show that the  $\theta$ -scheme (2.5) is unconditionally stable in the  $L^2$  norm if  $1/2 \leq \theta \leq 1$ , and stable under the CFL condition  $2(1 - 2\theta)\nu\Delta t \leq (\Delta x)^2$  if  $0 \leq \theta < 1/2$ .

**Exercise 2.2.5** Show that the 6-point scheme (2.6) is unconditionally stable in the  $L^2$  norm.

**Remark 2.2.19** Some authors use another definition of the stability, which is less restrictive than definition 2.2.8 but more complex. In this definition the scheme is called stable for the norm  $\| \cdot \|$  if for all time  $T > 0$  there exists a constant  $K(T) > 0$  independent of  $\Delta t$  and  $\Delta x$  such that

$$\|u^n\| \leq K(T)\|u^0\| \quad \text{for all } 0 \leq n \leq T/\Delta t,$$

whatever the initial data  $u^0$ . This new definition allows the solution to grow with time as is the case, for example, for the solution of the equation

$$\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} = cu \quad \text{for } (t, x) \in \mathbb{R}^+ \times \mathbb{R},$$

which, by changing the unknown  $v(t, x) = e^{-ct}u(t, x)$ , reduces to the heat equation (then the solution  $u$  grows exponentially in time). With such a definition of the stability, the Von Neumann stability condition becomes the inequality

$$|A(k)| \leq 1 + C\Delta t \quad \text{for all modes } k \in \mathbb{Z}.$$

For simplicity, we shall take the definition 2.2.8 of the stability. •

## 2.2.4 Convergence of the schemes

We now have all the tools to prove convergence of the finite differences schemes. The principal result of this section is the Lax theorem which shows that, for a linear scheme, **consistency and stability implies convergence**. The importance of this result far exceeds the finite difference method. For every numerical method (finite differences, finite elements, etc.) convergence is shown by combining two arguments: stability and consistency (their precise definitions change from one method to the other). From a practical point of view, the Lax theorem is very reassuring: if we use a consistent scheme (we can construct this generally) and we do not observe numerical oscillations (that is, it is stable), then the numerical solution is close to the exact solution (the scheme converges).

**Theorem 2.2.20 (Lax)** *Let  $u(t, x)$  be the sufficiently regular solution of the heat equation (2.1) (with the appropriate boundary conditions). Let  $u_j^n$  be the discrete numerical solution obtained by a finite difference scheme with the initial data  $u_j^0 = u_0(x_j)$ . We assume that the scheme is linear, two level, consistent, and stable for a norm  $\| \cdot \|$ . Then the scheme is convergent in the sense where*

$$\forall T > 0, \quad \lim_{\Delta t, \Delta x \rightarrow 0} \left( \sup_{t_n \leq T} \|e^n\| \right) = 0, \quad (2.20)$$

with  $e^n$  is the ‘error’ vector defined by its components  $e_j^n = u_j^n - u(t_n, x_j)$ .

Further, if the scheme has accuracy of order  $p$  in space and order  $q$  in time, then for all time  $T > 0$  there exists a constant  $C_T > 0$  such that

$$\sup_{t_n \leq T} \|e^n\| \leq C_T \left( (\Delta x)^p + (\Delta t)^q \right). \quad (2.21)$$

**Remark 2.2.21** We have not proved the existence and uniqueness of the solution of the heat equation (2.1) (with Dirichlet or periodic boundary conditions). For the moment, we therefore make the hypothesis of the existence and uniqueness of such a solution (as well as its regularity), but we see in Chapter 8 that this result is generally true. •

**Proof.** For simplicity, we assume that the boundary conditions are Dirichlet. The same proof is also true for periodic boundary conditions or for Neumann boundary conditions (assuming they are discretized with the same order of accuracy as the scheme). A two level linear scheme can be written in the condensed form (2.13), that is,

$$u^{n+1} = Au^n,$$

where  $A$  is the iteration matrix (square of size  $N$ ). Let  $u$  be the solution (assumed sufficiently regular) of the heat equation (2.1). We denote by  $\tilde{u}^n = (\tilde{u}_j^n)_{1 \leq j \leq N}$  with  $\tilde{u}_j^n = u(t_n, x_j)$ . As the scheme is consistent, there exists a vector  $\epsilon^n$  such that

$$\tilde{u}^{n+1} = A\tilde{u}^n + \Delta t \epsilon^n \quad \text{with} \quad \lim_{\Delta t, \Delta x \rightarrow 0} \|\epsilon^n\| = 0, \quad (2.22)$$

and the convergence of  $\epsilon^n$  is uniform for all time  $0 \leq t_n \leq T$ . If the scheme is accurate with order  $p$  in space and order  $q$  in time, then  $\|\epsilon^n\| \leq C((\Delta x)^p + (\Delta t)^q)$ . By setting  $e_j^n = \tilde{u}_j^n - u(t_n, x_j)$  we obtain by subtraction of (2.22) from (2.13)

$$e^{n+1} = Ae^n - \Delta t \epsilon^n$$

from which, by induction

$$e^n = A^n e^0 - \Delta t \sum_{k=1}^n A^{n-k} \epsilon^{k-1}. \quad (2.23)$$

Now, the stability of the scheme means that  $\|u^n\| = \|A^n u^0\| \leq K \|u^0\|$  for all initial data, that is,  $\|A^n\| \leq K$  where the constant  $K$  does not depend on  $n$ . On the other hand,  $e^0 = 0$ , therefore (2.23) gives

$$\|e^n\| \leq \Delta t \sum_{k=1}^n \|A^{n-k}\| \|\epsilon^{k-1}\| \leq \Delta t n K C \left( (\Delta x)^p + (\Delta t)^q \right),$$

which gives the inequality (2.21) with the constant  $C_T = TKC$ . The proof of (2.20) is similar. □

**Remark 2.2.22** The Lax theorem 2.2.20 is in fact valid for all linear partial differential equations. It has a converse in the sense that if a two level linear consistent scheme is convergent then it must be stable. We remark that the rate of convergence in (2.21) is exactly the accuracy of the scheme. Finally, it is good to note that the estimate (2.21) is only valid on a bounded time interval  $[0, T]$  but it is independent of the number of points of discretization  $N$ . •

### 2.2.5 Multilevel schemes

Up until now we have mainly analysed two level schemes, that is, the schemes which relate the values of  $u^{n+1}$  only to those of  $u^n$ . We can easily envisage multilevel schemes, and in particular we have already introduced some three level schemes where  $u^{n+1}$  depends on  $u^n$  and  $u^{n-1}$  (like the Richardson, DuFort–Frankel, or Gear schemes). We shall now study how the previous results generalize to multilevel schemes (we limit ourselves for sake of clarity to three level schemes).

The definition 2.2.8 of the stability of a scheme is independent of its number of levels. However, the interpretation of the stability in terms of the iteration matrix is a little more complicated for a three level linear scheme. Indeed,  $u^{n+1}$  depends linearly on  $u^n$  and  $u^{n-1}$ , therefore, we cannot write the relation (2.13). But, if we set

$$U^n = \begin{pmatrix} u^n \\ u^{n-1} \end{pmatrix}, \quad (2.24)$$

then there exist two matrices of order  $N$ ,  $A_1$ , and  $A_2$ , such that

$$U^{n+1} = AU^n = \begin{pmatrix} A_1 & A_2 \\ I & 0 \end{pmatrix} U^n, \quad (2.25)$$

where the iteration matrix  $A$  is therefore of size  $2N$ . As before,  $U^n = A^n U^1$  and the stability is equivalent to

$$\|A^n\| = \sup_{U^1 \in \mathbb{R}^{2N}, U^1 \neq 0} \frac{\|A^n U^1\|}{\|U^1\|} \leq K \quad \forall n \geq 1.$$

In the same way Fourier analysis extends to three level schemes thanks to vector notation (2.24). As an example, we prove a result presented in Chapter 1.

**Lemma 2.2.23** *The centred scheme (1.28) is unstable in the  $L^2$  norm.*

**Proof.** With the usual notation the scheme (1.28) is written, for  $x \in [0, 1]$ ,

$$\frac{u^{n+1}(x) - u^{n-1}(x)}{2\Delta t} + \nu \frac{-u^n(x - \Delta x) + 2u^n(x) - u^n(x + \Delta x)}{(\Delta x)^2} = 0,$$

and by application of the Fourier transform

$$\hat{u}^{n+1}(k) + \frac{8\nu\Delta t}{(\Delta x)^2} (\sin(\pi k \Delta x))^2 \hat{u}^n(k) - \hat{u}^{n-1}(k) = 0.$$

In other words,

$$\hat{U}^{n+1}(k) = \begin{pmatrix} \hat{u}^{n+1}(k) \\ \hat{u}^n(k) \end{pmatrix} = \begin{pmatrix} -\frac{8\nu\Delta t}{(\Delta x)^2} (\sin(\pi k \Delta x))^2 & 1 \\ 1 & 0 \end{pmatrix} \hat{U}^n(k) = A(k) \hat{U}^n(k),$$

and  $\hat{U}^{n+1}(k) = A(k)^n \hat{U}^1(k)$ . Here,  $A(k)$  is a matrix of order 2 which for two level schemes was a scalar. For  $k \in \mathbb{Z}$ , the vector  $\hat{U}^n(k)$ , and therefore the Fourier coefficient  $\hat{u}^n(k)$ , is bounded as  $n$  tends to infinity if and only if the amplification matrix satisfies

$$\|A(k)^n\|_2 = \sup_{U \in \mathbb{R}^2, U \neq 0} \frac{\|A(k)^n U\|_2}{\|U\|_2} \leq K \quad \forall n \geq 1, \quad (2.26)$$

where  $\|U\|_2$  is the Euclidean norm in  $\mathbb{R}^2$ . Consequently, if the inequality (2.26) is true for arbitrary Fourier modes  $k \in \mathbb{Z}$ , by the Plancherel formula we deduce

$$\|u^n\|_2^2 = \sum_{k \in \mathbb{Z}} |\hat{u}^n(k)|^2 \leq K \sum_{k \in \mathbb{Z}} (|\hat{u}^0(k)|^2 + |\hat{u}^1(k)|^2) = \|u^0\|_2^2 + \|u^1\|_2^2,$$

that is, the  $L^2$  stability of the scheme. Conversely, if there exists  $k_0$  such that  $\|A(k_0)^n\|$  is not bounded as  $n$  tends to infinity, then by a suitable choice of the initial data with a single mode  $\hat{u}^0(k_0)$  (like  $\hat{u}^1(k_0)$ ), we obtain the  $L^2$  instability of the scheme.

As the amplification matrix  $A(k)$  is real symmetric, we have the property  $\|A(k)\|_2 = \rho(A(k))$  and  $\|A(k)^n\|_2 = \|A(k)\|_2^n$ , where  $\rho(M)$  denotes the spectral radius of the matrix  $M$  (see lemma 13.1.6). Therefore, the inequality (2.26) is satisfied if and only if  $\rho(A(k)) \leq 1$ . The eigenvalues of  $A(k)$  are the roots of the second degree polynomial

$$\lambda^2 + \frac{8\nu\Delta t}{(\Delta x)^2} (\sin(\pi k \Delta x))^2 \lambda - 1 = 0$$

which always has real distinct roots with product  $-1$ . Consequently, one of the roots is (strictly) greater than 1 in modulus, and thus  $\rho(A(k)) > 1$ . Therefore, the centred scheme is unconditionally unstable in the  $L^2$  norm.  $\square$

**Remark 2.2.24** The Fourier analysis which we have used in the proof of lemma 2.2.23 is a little more complicated in the case of multilevel schemes than in two level schemes (see remark 2.2.18). When we put Fourier modes into the scheme, we obtain

$$\begin{pmatrix} u_j^{n+1} \\ u_j^n \end{pmatrix} = A(k)^n \begin{pmatrix} u_j^1 \\ u_j^0 \end{pmatrix} \exp(2i\pi k x_j)$$

where  $A(k)$  is from now on an amplification **matrix** (and no longer a scalar factor). We write: **Von Neumann stability condition** to mean the condition

$$\rho(A(k)) \leq 1 \quad \text{for all modes } k \in \mathbb{Z}, \quad (2.27)$$

where  $\rho(A(k))$  is the spectral radius of the matrix  $A(k)$ . Since for an arbitrary matrix  $B$  we have

$$\|B\| \geq \rho(B) \quad \text{and} \quad \|B^n\| \geq \rho(B)^n,$$

it is clear that the Von Neumann stability condition is a **necessary condition** for  $L^2$  stability of the scheme (therefore of its convergence). If the matrix  $A(k)$  is normal,

it satisfies  $\|A(k)\|_2 = \rho(A(k))$  and  $\|A(k)^n\|_2 = \|A(k)\|_2^n$  (see lemma 13.1.6), therefore the Von Neumann condition (2.27) is necessary and sufficient (we had the ‘luck’ in the proof of lemma 2.2.23 to be in this favourable case). However, if  $A(k)$  is not normal, then in general, the Von Neumann stability condition is **not sufficient** and we must make a more delicate analysis of  $A(k)$  (and in particular of its diagonalization). •

**Remark 2.2.25** The Lax theorem 2.2.20 generalizes without difficulty to multilevel schemes if we choose the  $L^2$  norm. The method of proof is unchanged: it uses Fourier analysis and vector notation (2.24). •

**Remark 2.2.26** Everything which we have said about the stability and the convergence of multilevel schemes generalizes immediately to schemes for systems of equations. In this case, we must also write a vectorial version of the recurrence relation (2.25) and of the amplification matrix (instead of a scalar factor). •

**Exercise 2.2.6** Show that the Gear scheme (2.8) is unconditionally stable and therefore convergent in the  $L^2$  norm.

**Exercise 2.2.7** Show that the DuFort–Frankel scheme (2.7) is stable in the  $L^2$  norm and therefore convergent, if the ratio  $\Delta t/(\Delta x)^2$  remains bounded as we let  $\Delta t$  and  $\Delta x$  tend to 0.

## 2.2.6 The multidimensional case

The finite difference method extends without difficulty to problems in several space dimensions. Let us consider, for example, the heat equation in two space dimensions (the case of there or more space dimensions is not more complicated, at least in theory) in the rectangular domain  $\Omega = (0, 1) \times (0, L)$  with the Dirichlet boundary conditions

$$\begin{cases} \frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} - \nu \frac{\partial^2 u}{\partial y^2} = 0 & \text{for } (x, y, t) \in \Omega \times \mathbb{R}_*^+ \\ u(t = 0, x, y) = u_0(x, y) & \text{for } (x, y) \in \Omega \\ u(t, x, y) = 0 & \text{for } t \in \mathbb{R}_*^+, (x, y) \in \partial\Omega. \end{cases} \quad (2.28)$$

To discretize the domain  $\Omega$ , we introduce two space steps  $\Delta x = 1/(N_x + 1) > 0$  and  $\Delta y = L/(N_y + 1) > 0$  (with  $N_x$  and  $N_y$  being two positive integers). With the time step  $\Delta t > 0$ , we define the nodes of a regular mesh (see Figure 2.1)

$$(t_n, x_j, y_k) = (n\Delta t, j\Delta x, k\Delta y) \quad \text{for } n \geq 0, 0 \leq j \leq N_x + 1, 0 \leq k \leq N_y + 1.$$

We denote by  $u_{j,k}^n$  the value of an approximate discrete solution at the point  $(t_n, x_j, y_k)$ , and  $u(t, x, y)$  the exact solution of (2.28).

The Dirichlet boundary conditions are expressed, for  $n > 0$ , as

$$u_{0,k}^n = u_{N_x+1,k}^n = 0, \quad \forall k, \quad \text{and} \quad u_{j,0}^n = u_{j,N_y+1}^n = 0, \quad \forall j.$$

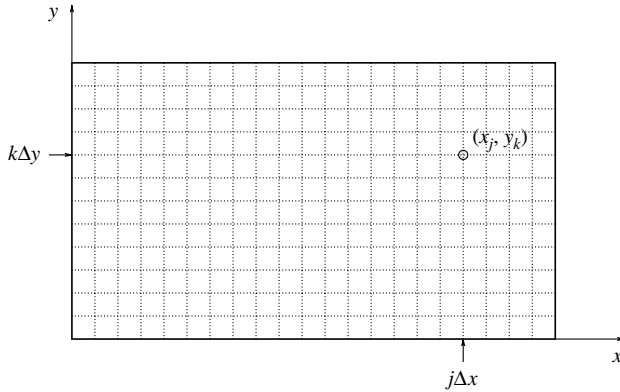


Figure 2.1. Rectangular finite difference mesh.

The initial data is discretized by

$$u_{j,k}^0 = u_0(x_j, y_k) \quad \forall j, k.$$

The generalization to the two-dimensional case of the **explicit scheme** is obvious

$$\frac{u_{j,k}^{n+1} - u_{j,k}^n}{\Delta t} + \nu \frac{-u_{j-1,k}^n + 2u_{j,k}^n - u_{j+1,k}^n}{(\Delta x)^2} + \nu \frac{-u_{j,k-1}^n + 2u_{j,k}^n - u_{j,k+1}^n}{(\Delta y)^2} = 0 \quad (2.29)$$

for  $n \geq 0$ ,  $j \in \{1, \dots, N_x\}$  and  $k \in \{1, \dots, N_y\}$ . The only notable difference with the one-dimensional case is the extra severity of the CFL condition.

**Exercise 2.2.8** Show that the explicit scheme (2.29) is stable in the  $L^\infty$  norm (and that it satisfies the maximum principle) under the CFL condition

$$\frac{\nu \Delta t}{(\Delta x)^2} + \frac{\nu \Delta t}{(\Delta y)^2} \leq \frac{1}{2}.$$

We illustrate the explicit scheme (2.29) (to which we add a convection term) by Figure 2.2 which represents convection–diffusion of a ‘hump’ (the coefficient of diffusion is 0.01 and the velocity  $(1, 0)$ ).

Likewise, we have the **implicit scheme**

$$\frac{u_{j,k}^{n+1} - u_{j,k}^n}{\Delta t} + \nu \frac{-u_{j-1,k}^{n+1} + 2u_{j,k}^{n+1} - u_{j+1,k}^{n+1}}{(\Delta x)^2} + \nu \frac{-u_{j,k-1}^{n+1} + 2u_{j,k}^{n+1} - u_{j,k+1}^{n+1}}{(\Delta y)^2} = 0. \quad (2.30)$$

We remark that the implicit scheme needs, to calculate  $u^{n+1}$  as a function of  $u^n$ , the solution of a linear system significantly more complicated than that in one space



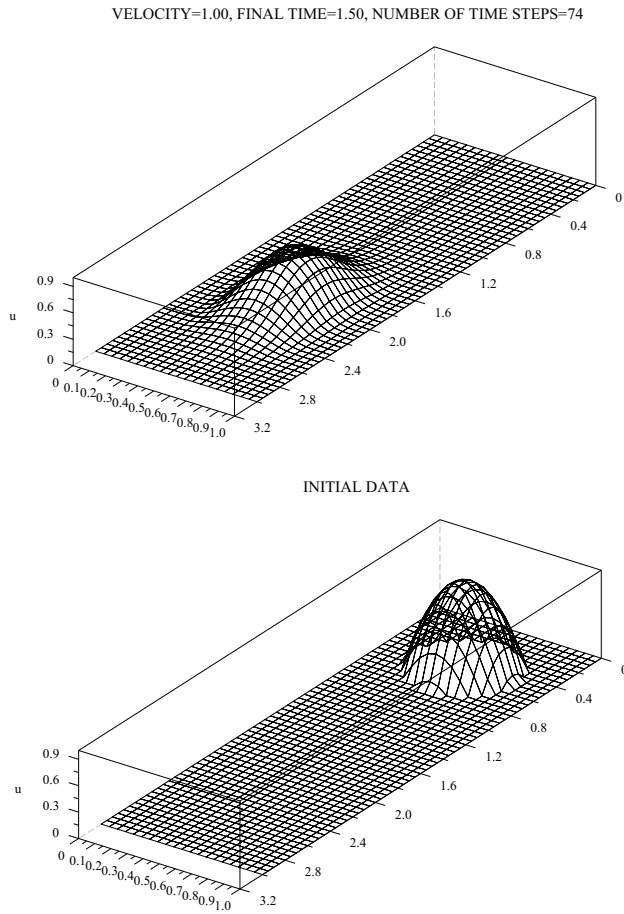


Figure 2.2. Explicit scheme for the convection–diffusion equation in two dimensions: initial data (top) and solution (bottom).

dimension (the situation will be even worse in three dimensions). Recall that in one dimension, it is sufficient to invert a tridiagonal matrix. We shall see that in two dimensions the matrix has a less simple structure. The discrete unknown  $u_{j,k}^n$  is indexed by two integers  $j$  and  $k$ , but in practice we use only one index to store  $u^n$  in the form of a vector in the computer. A (simple and efficient) way of putting the unknowns  $u_{j,k}^n$  into a single vector is to write

$$u^n = (u_{1,1}^n, \dots, u_{1,N_y}^n, u_{2,1}^n, \dots, u_{2,N_y}^n, \dots, u_{N_x,1}^n, \dots, u_{N_x,N_y}^n).$$

Note that we have arranged the unknowns ‘column by column’, but we could equally well have done it ‘row by row’ by using the  $j$  index first instead of the  $k$  index ( $N_x$  is the number of columns and  $N_y$  the number of rows). With this convention, the implicit scheme (2.30) requires the inversion of the matrix, which is ‘block’ symmetric tridiagonal,

$$M = \begin{pmatrix} D_1 & E_1 & & & 0 \\ E_1 & D_2 & E_2 & & \\ & \ddots & \ddots & \ddots & \\ & & E_{N_x-2} & D_{N_x-1} & E_{N_x-1} \\ 0 & & & E_{N_x-1} & D_{N_x} \end{pmatrix}$$

where the diagonal blocks  $D_j$  are square matrices of dimension  $N_y$

$$D_j = \begin{pmatrix} 1 + 2(c_y + c_x) & -c_y & & & 0 \\ -c_y & 1 + 2(c_y + c_x) & -c_y & & \\ & \ddots & \ddots & \ddots & \\ & & -c_y & 1 + 2(c_y + c_x) & -c_y \\ 0 & & & -c_y & 1 + 2(c_y + c_x) \end{pmatrix}$$

with  $c_x = \nu \Delta t / (\Delta x)^2$  and  $c_y = \nu \Delta t / (\Delta y)^2$ , and the extra-diagonal blocks  $E_j = (E_j)^*$  are square matrices of dimension  $N_y$

$$E_j = \begin{pmatrix} -c_x & 0 & & & 0 \\ 0 & -c_x & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & 0 & -c_x & 0 \\ 0 & & & 0 & -c_x \end{pmatrix}.$$

In summary, the matrix  $M$  is symmetric and pentadiagonal. However, the five diagonals are not contiguous, this implies a considerable extra cost to solve a linear system associated with  $M$  (see the appendix on numerical linear algebra and particularly the remarks 13.1.21 and 13.1.41). The situation will be even worse in three dimensions.

**Exercise 2.2.9** Show that the Peaceman–Rachford scheme

$$\begin{aligned} \frac{u_{j,k}^{n+1/2} - u_{j,k}^n}{\Delta t} + \nu \frac{-u_{j-1,k}^{n+1/2} + 2u_{j,k}^{n+1/2} - u_{j+1,k}^{n+1/2}}{2(\Delta x)^2} + \nu \frac{-u_{j,k-1}^n + 2u_{j,k}^n - u_{j,k+1}^n}{2(\Delta y)^2} &= 0 \\ \frac{u_{j,k}^{n+1} - u_{j,k}^{n+1/2}}{\Delta t} + \nu \frac{-u_{j-1,k}^{n+1/2} + 2u_{j,k}^{n+1/2} - u_{j+1,k}^{n+1/2}}{2(\Delta x)^2} + \nu \frac{-u_{j,k-1}^{n+1} + 2u_{j,k}^{n+1} - u_{j,k+1}^{n+1}}{2(\Delta y)^2} &= 0. \end{aligned}$$

has accuracy of order 2 in space and time and is unconditionally stable in the  $L^2$  norm (for periodic boundary conditions in each direction).

Because of the heightened cost of calculation, we often replace the implicit scheme by a generalization to several space dimensions of the one-dimensional scheme, obtained by a technique called **alternating directions**, also called operator splitting, or **splitting**. The idea is, instead of solving the two-dimensional equation (2.28), we solve alternatively the two one-dimensional equations

$$\frac{\partial u}{\partial t} - 2\nu \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{and} \quad \frac{\partial u}{\partial t} - 2\nu \frac{\partial^2 u}{\partial y^2} = 0$$

whose average again gives (2.28). For example, by using a Crank–Nicolson scheme in each direction for a half time step  $\Delta t/2$ , we obtain an **alternating direction scheme**

$$\begin{aligned} \frac{u_{j,k}^{n+1/2} - u_{j,k}^n}{\Delta t} + \nu \frac{-u_{j-1,k}^{n+1/2} + 2u_{j,k}^{n+1/2} - u_{j+1,k}^{n+1/2}}{2(\Delta x)^2} + \nu \frac{-u_{j,k-1}^n + 2u_{j,k}^n - u_{j,k+1}^n}{2(\Delta x)^2} &= 0 \\ \frac{u_{j,k}^{n+1} - u_{j,k}^{n+1/2}}{\Delta t} + \nu \frac{-u_{j,k-1}^{n+1} + 2u_{j,k}^{n+1} - u_{j,k+1}^{n+1}}{2(\Delta y)^2} + \nu \frac{-u_{j,k-1}^{n+1/2} + 2u_{j,k}^{n+1/2} - u_{j,k+1}^{n+1/2}}{2(\Delta y)^2} &= 0 \end{aligned} \tag{2.31}$$

The advantage of this type of scheme is that it is enough, at each half time step, to invert a ‘one-dimensional’ tridiagonal matrix (therefore an inexpensive calculation). In three dimensions, it is enough to take three one-third time steps and the properties of the scheme are unchanged. This scheme is not only stable but also consistent with the two-dimensional equation (2.28).

**Exercise 2.2.10** Show that the alternating direction scheme (2.31) has accuracy of order 2 in space and time and is unconditionally stable in the  $L^2$  norm (for periodic boundary conditions in each direction).

Let us conclude this section with some practical considerations for the finite difference method. Its principal advantage is its simplicity as well as its computational implementation. However, it has a certain number of defects which, for many complex problems, leads us to prefer other methods such as the finite element method (see Chapters 6 and 8). One of the principal limitations of the method is that it only works

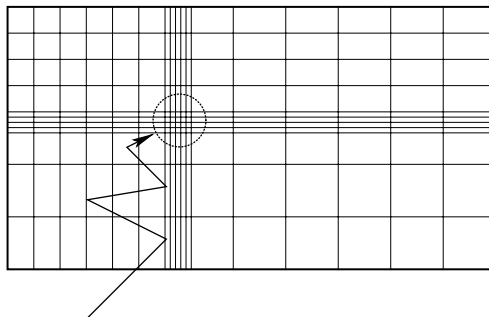


Figure 2.3. Refinement of a finite difference mesh: the encircled zone is that where we want more accuracy.

for regular, or **rectangular**, meshes. It is not always easy to discretize an arbitrary space domain by rectangular meshes! Additionally, it is not possible to locally refine the mesh to have better accuracy at a particular point of the domain. It is possible to vary the space step in each direction but this variation is uniform in perpendicular directions ( $\Delta x$  and  $\Delta y$  can change along the  $x$  and  $y$  axes, respectively, but this variation is uniform in orthogonal directions; see Figure 2.3). Such a refinement of a finite difference mesh therefore has effects far outside the zone of interest. Moreover, the theory and practice of the finite differences become much more complicated when the coefficients in the partial differential equations are variables and when the problems are nonlinear.

## 2.3 Other models

### 2.3.1 Advection equation

We consider the advection equation in one space dimension in the bounded domain  $(0, 1)$  with a constant velocity  $V > 0$  and with the periodic boundary conditions

$$\begin{cases} \frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} = 0 & \text{for } (x, t) \in (0, 1) \times \mathbb{R}_+^+ \\ u(t, x + 1) = u(t, x) & \text{for } (x, t) \in (0, 1) \times \mathbb{R}_+^+ \\ u(0, x) = u_0(x) & \text{for } x \in (0, 1). \end{cases} \quad (2.32)$$

We always discretize space with a step  $\Delta x = 1/(N + 1) > 0$  ( $N$  a positive integer) and the time with  $\Delta t > 0$ , and we denote by  $(t_n, x_j) = (n\Delta t, j\Delta x)$  for  $n \geq 0, j \in \{0, 1, \dots, N + 1\}$ ,  $u_j^n$  the value of an approximate discrete solution at the point  $(t_n, x_j)$ , and  $u(t, x)$  the exact solution of (2.32). The periodic boundary conditions lead to

equations  $u_0^n = u_{N+1}^n$  for all  $n \geq 0$ , and more generally  $u_j^n = u_{N+1+j}^n$ . Consequently, the discrete unknown at each time step is a vector  $u^n = (u_j^n)_{0 \leq j \leq N} \in \mathbb{R}^{N+1}$ . We give some possible schemes for the advection equation (2.32). In Chapter 1 we have already noted the bad numerical behaviour of the **explicit centred scheme**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0 \quad (2.33)$$

for  $n \geq 0$  and  $j \in \{0, \dots, N\}$ . The unstable character of this scheme is confirmed by the following lemma.

**Lemma 2.3.1** *The explicit centred scheme (2.33) is consistent with the advection equation (2.32), accurate with order 1 in time and 2 in space, but unconditionally unstable in the  $L^2$  norm.*

**Proof.** With the help of a Taylor expansion around the point  $(t_n, x_j)$ , we easily see that the scheme is consistent, accurate with order 1 in time and 2 in space. By Fourier analysis, we study the  $L^2$  stability. With the notation of Section 2.2.3, the Fourier components  $\hat{u}^n(k)$  of  $u^n$  satisfy

$$\hat{u}^{n+1}(k) = \left(1 - i \frac{V\Delta t}{\Delta x} \sin(2\pi k\Delta x)\right) \hat{u}^n(k) = A(k) \hat{u}^n(k).$$

We see that the amplification factor is always greater than 1,

$$|A(k)|^2 = 1 + \left(\frac{V\Delta t}{\Delta x} \sin(2\pi k\Delta x)\right)^2 \geq 1,$$

with strict inequality when  $2k\Delta x$  is not an integer. Therefore the scheme is unstable.  $\square$

We can write an implicit version of the preceding scheme which is stable: it is the **implicit centred scheme**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\Delta x} = 0. \quad (2.34)$$

**Exercise 2.3.1** Show that the implicit centred scheme (2.34) is consistent with the advection equation (2.32), accurate with order 1 in time and 2 in space, unconditionally stable in the  $L^2$  norm, and therefore convergent.

If we absolutely must stay centred and explicit, the **Lax–Friedrichs scheme**

$$\frac{2u_j^{n+1} - u_{j+1}^n - u_{j-1}^n}{2\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0 \quad (2.35)$$

is a scheme which is simple, robust but not very accurate.

**Lemma 2.3.2** *The Lax–Friedrichs scheme (2.35) is stable in the  $L^2$  norm under the CFL condition*

$$|V|\Delta t \leq \Delta x.$$

*If the ratio  $\Delta t/\Delta x$  is held constant as  $\Delta t$  and  $\Delta x$  tend to zero, it is consistent with the advection equation (2.32) and accurate with order 1 in space and time. Consequently, it is conditionally convergent.*

**Proof.** By Fourier analysis we have

$$\hat{u}^{n+1}(k) = \left( \cos(2\pi k \Delta x) - i \frac{V \Delta t}{\Delta x} \sin(2\pi k \Delta x) \right) \hat{u}^n(k) = A(k) \hat{u}^n(k).$$

The modulus of the amplification factor is given by

$$|A(k)|^2 = \cos^2(2\pi k \Delta x) + \left( \frac{V \Delta t}{\Delta x} \right)^2 \sin^2(2\pi k \Delta x).$$

We see, therefore, that  $|A(k)| \leq 1$  for all  $k$  if the condition  $|V|\Delta t \leq \Delta x$  is satisfied, while if not there exist unstable modes  $k$  such that  $|A(k)| > 1$ . The scheme is therefore conditionally stable. To study the consistency, we make a Taylor expansion around  $(t_n, x_j)$  for the solution  $u$ :

$$\begin{aligned} & \frac{2u(t_{n+1}, x_j) - u(t_n, x_{j+1}) - u(t_n, x_{j-1}))}{2\Delta t} + V \frac{u(t_n, x_{j+1}) - u(t_n, x_{j-1}))}{2\Delta x} = \\ & (u_t + V u_x)(t_n, x_j) - \frac{(\Delta x)^2}{2\Delta t} \left( 1 - \frac{(V \Delta t)^2}{(\Delta x)^2} \right) u_{xx}(t_n, x_j) + \mathcal{O}\left((\Delta x)^2 + \frac{(\Delta x)^4}{\Delta t}\right). \end{aligned} \quad (2.36)$$

Since the truncation error contains a term in  $\mathcal{O}((\Delta x)^2/\Delta t)$ , the scheme is not consistent if  $\Delta t$  tends to zero more quickly than  $(\Delta x)^2$ . Conversely, it is consistent and is accurate with order 1 if the ratio  $\Delta t/\Delta x$  is constant. To obtain the convergence we recall the proof of the Lax Theorem 2.2.20. The error  $e^n$  is always bounded by the truncation error, and therefore

$$\|e^n\| \leq \Delta t n K C \left( \frac{(\Delta x)^2}{\Delta t} + \Delta t \right).$$

If we keep the ratio  $\Delta x/\Delta t$  fixed, the error is therefore bounded by a constant times  $\Delta t$  which tends to zero, from which we have the convergence.  $\square$

**Remark 2.3.3** The Lax–Friedrichs scheme is not (in the strict sense of the definition 2.2.4) consistent. Nevertheless, it is conditionally consistent and convergent. We must, however, pay attention to the fact that if we take a much smaller time step  $\Delta t$  than is permitted by the CFL stability condition, the convergence will be very slow. In practice, the Lax–Friedrichs scheme is not recommended.  $\bullet$

An explicit centred scheme which is more accurate is **Lax–Wendroff scheme**

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} - \left( \frac{V^2 \Delta t}{2} \right) \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{(\Delta x)^2} = 0. \quad (2.37)$$

The derivation is not immediate, we present it in detail. We start by writing an expansion of order 2 in time of the exact solution

$$u(t_{n+1}, x_j) = u(t_n, x_j) + (\Delta t)u_t(t_n, x_j) + \frac{(\Delta t)^2}{2}u_{tt}(t_n, x_j) + \mathcal{O}\left((\Delta t)^3\right).$$

By using the advection equation we replace the time derivatives by space derivatives

$$u(t_{n+1}, x_j) = u(t_n, x_j) - (V\Delta t)u_x(t_n, x_j) + \frac{(V\Delta t)^2}{2}u_{xx}(t_n, x_j) + \mathcal{O}\left((\Delta t)^3\right).$$

Finally, we replace the space derivatives by a centred formula of order 2

$$\begin{aligned} u(t_{n+1}, x_j) &= u(t_n, x_j) - V\Delta t \frac{u(t_n, x_{j+1}) - u(t_n, x_{j-1}))}{2\Delta x} \\ &\quad + \frac{(V\Delta t)^2}{2} \frac{u(t_n, x_{j+1}) - 2u(t_n, x_j) + u(t_n, x_{j-1}))}{(\Delta x)^2} + \mathcal{O}\left((\Delta t)^3 + \Delta t(\Delta x)^2\right). \end{aligned}$$

We recover the Lax–Wendroff scheme by neglecting the third order terms and replacing  $u(t_n, x_j)$  by  $u_j^n$ . We remark that, compared to the previous schemes, we have ‘simultaneously’ discretized the space and time derivatives of the advection equation. By design, the Lax–Wendroff scheme is accurate with order 2 in time and in space. We can show that it does not satisfy the discrete maximum principle (see exercise 2.3.3). Conversely, it is stable in the  $L^2$  norm and therefore convergent under the CFL condition  $|V|\Delta t \leq \Delta x$ .

**Exercise 2.3.2** Show that the Lax–Wendroff scheme is stable and convergent in the  $L^2$  norm if  $|V|\Delta t \leq \Delta x$ .

**Exercise 2.3.3** Show that the Lax–Friedrichs scheme satisfies the discrete maximum principle if the CFL condition  $|V|\Delta t \leq \Delta x$  is satisfied, while the Lax–Wendroff scheme does not satisfy it except if  $V\Delta t/\Delta x$  is  $-1, 0$ , or  $1$ .

**Exercise 2.3.4** Show that the Lax–Wendroff scheme (2.37) is the only scheme which is accurate with order 2 in space and time of the type

$$u_j^{n+1} = \alpha u_{j-1}^n + \beta u_j^n + \gamma u_{j+1}^n,$$

where  $\alpha, \beta, \gamma$  depend only on  $V\Delta t/\Delta x$ .

As we have already seen in Chapter 1, a fundamental idea to obtain ‘good’ schemes for the advection equation (2.32) is **upwinding**. We give the general form of the **upwinded scheme**

$$\begin{aligned} \frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_j^n - u_{j-1}^n}{\Delta x} &= 0 & \text{if } V > 0 \\ \frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_j^n}{\Delta x} &= 0 & \text{if } V < 0. \end{aligned} \quad (2.38)$$

We have already seen in Chapter 1 that the upwinded scheme is stable in the  $L^\infty$  norm if the CFL condition,  $|V|\Delta t \leq \Delta x$ , is satisfied. As it is consistent and accurate with order 1 in space and time, it converges in the  $L^\infty$  norm from the Lax theorem. The same result is true in the  $L^2$  norm with the same CFL condition.

**Exercise 2.3.5** Show that the explicit upwinded scheme (2.38) is consistent with the advection equation (2.32), accurate with order 1 in space and time, stable and convergent in the  $L^2$  norm if the CFL condition  $|V|\Delta t \leq \Delta x$  is satisfied.

**Remark 2.3.4** For nonlinear problems (where the velocity  $V$  itself depends on the unknown  $u$ ), and particularly for fluid flow models, the upwinded scheme is clearly superior to the others. It is the source of many generalizations, much more complex than the original (see [23]). In particular, even though the original scheme is only of order 1, it has variants of order 2. •

Scheme	Stability	Truncation error
Explicit centred (2.33)	Unstable	$\mathcal{O}(\Delta t + (\Delta x)^2)$
Implicit centred (2.34)	$L^2$ stable	$\mathcal{O}(\Delta t + (\Delta x)^2)$
Lax–Friedrichs (2.35)	$L^2$ and $L^\infty$ stable for the CFL condition $ V \Delta t \leq \Delta x$	$\mathcal{O}(\Delta t + (\Delta x)^2/\Delta t)$
Lax–Wendroff (2.37)	$L^2$ stable for the CFL condition $ V \Delta t \leq \Delta x$	$\mathcal{O}((\Delta t)^2 + (\Delta x)^2)$
Upwinded (2.38)	$L^2$ and $L^\infty$ stable for the CFL condition $ V \Delta t \leq \Delta x$	$\mathcal{O}(\Delta t + \Delta x)$

Table 2.2. Summary of properties of various schemes for the advection equation

To compare these various schemes (see Table 2.2) from a practical viewpoint, a pertinent (though formal) concept is that of the equivalent equation.

**Definition 2.3.5** We call the **equivalent equation** of a scheme the equation obtained by adding the principal part (that is, the term with dominant order) of the truncation error to the model studied.



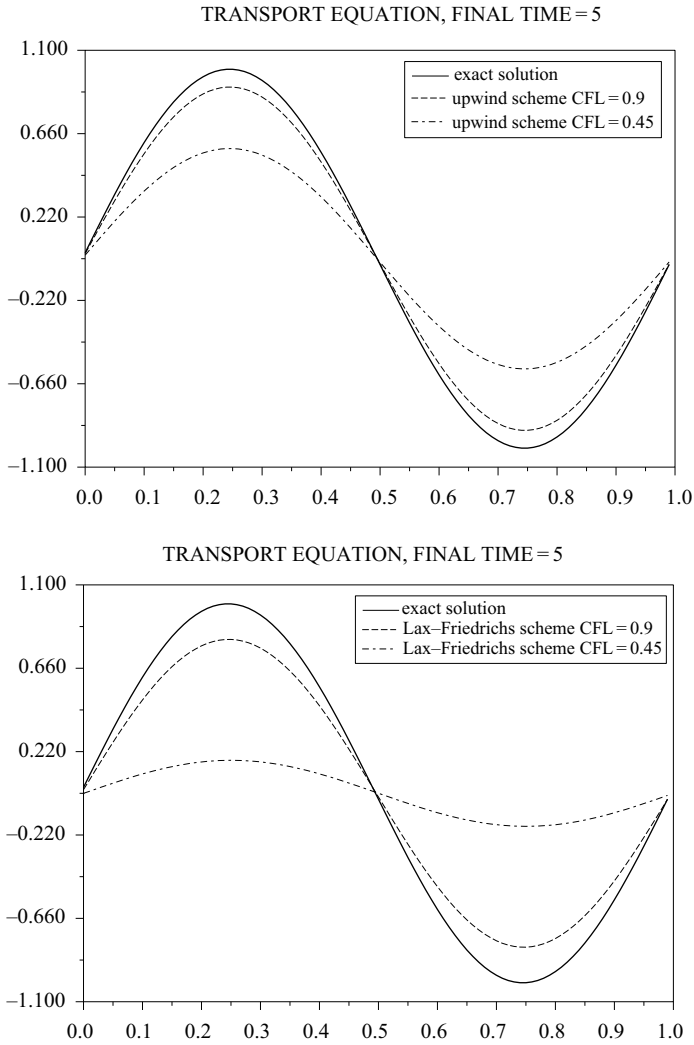


Figure 2.4. Influence of the CFL condition on the numerical diffusion of the Lax–Friedrichs scheme (top) and of the upwind scheme (bottom).

All of the schemes which we have seen are consistent. However, if we add the principal part of the truncation error of a scheme to the equation, then this scheme is not only consistent with this new ‘equivalent’ equation, but is also strictly more accurate for this equivalent equation. In other words, the scheme is ‘more consistent’ with the equivalent equation than with the original equation. Let us take the example of the Lax–Friedrichs scheme (2.35) for the advection equation: from (2.36), the principal part of its truncation error is  $-\frac{(\Delta x)^2}{2\Delta t} \left(1 - \frac{(V\Delta t)^2}{(\Delta x)^2}\right) u_{xx}$ . Consequently, the equivalent equation of the Lax–Friedrichs scheme is

$$\frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} - \nu \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{with } \nu = \frac{(\Delta x)^2}{2\Delta t} \left(1 - \frac{(V\Delta t)^2}{(\Delta x)^2}\right). \quad (2.39)$$

This equivalent equation will give us invaluable information on numerical behaviour of the scheme. Indeed, the Lax–Friedrichs scheme is a good approximation (of order 2) of the convection–diffusion equation (2.39) where the coefficient of diffusion  $\nu$  is small (even zero if the CFL condition is exactly satisfied, that is,  $\Delta x = |V|\Delta t$ ). We remark that if the time step is taken to be very small, the coefficient of diffusion  $\nu$  may be very large and the scheme is bad as it is too weighted to the diffusion (see Figure 2.4). The coefficient of diffusion  $\nu$  of the equivalent equation is called **numerical diffusion**. If it is large, we say that the scheme is **diffusive** (or dissipative). The typical behaviour of a diffusive scheme is its tendency to artificially spread out the initial data in the course of time. The schemes which are too diffusive are therefore ‘bad’ schemes.

**Exercise 2.3.6** Show that the equivalent equation of the upwinded scheme (2.38) is

$$\frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} - \frac{|V|}{2} (\Delta x - |V|\Delta t) \frac{\partial^2 u}{\partial x^2} = 0.$$

The upwinded scheme is also diffusive (except if the CFL condition is exactly satisfied, that is,  $\Delta x = |V|\Delta t$ ). In any case, the diffusion coefficient of the equivalent equation does not tend to infinity as the time step tends to zero (for  $\Delta x$  fixed), which is a clear improvement with respect to the Lax–Friedrichs scheme (see Figure 2.4). This numerical diffusion effect is illustrated by the Figure 2.4 where we solve the advection equation on an interval of length 1 with periodic boundary conditions, sinusoidal initial data, space step  $\Delta x = 0.01$ , velocity  $V = 1$  and final time  $T = 5$ . We compare two values of the time step  $\Delta t = 0.9\Delta x$  and  $\Delta t = 0.45\Delta x$ .

**Exercise 2.3.7** Show that the equivalent equation of the Lax–Wendroff scheme (2.37) is

$$\frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} + \frac{V(\Delta x)^2}{6} \left(1 - \frac{(V\Delta t)^2}{(\Delta x)^2}\right) \frac{\partial^3 u}{\partial x^3} = 0.$$

As the Lax–Wendroff scheme is accurate with order 2, the equivalent equation does not contain a diffusion term but a third order term, called **dispersive**. Let us remark that the coefficient of this dispersive term is of much smaller order than the coefficient

of diffusion of the equivalent equations for the diffusive schemes. This is why this dispersive effect can only, in general, be seen on a nondiffusive scheme. The typical behaviour of a dispersive scheme is that it produces oscillations when the solution is discontinuous (see Figure 2.5). In effect, the dispersive term modifies the velocity of propagation of the plane waves or Fourier modes of the solution (particularly of these modes with high frequency), whereas a diffusive term only attenuates its amplitude (see Exercise 2.3.8).

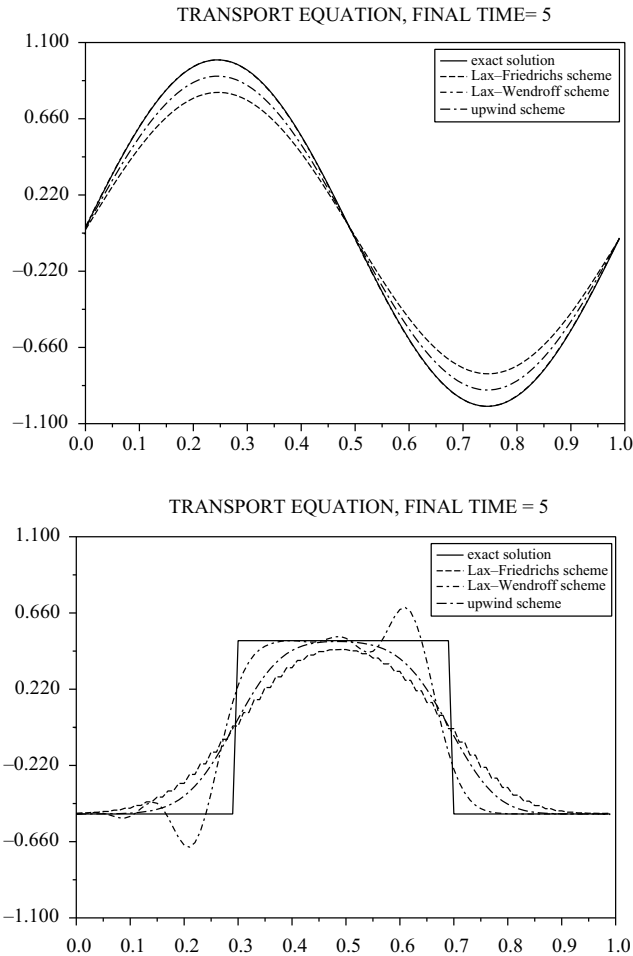


Figure 2.5. Comparison of the Lax–Friedrichs, Lax–Wendroff, and upwind schemes for sinusoidal initial data (top) and square wave (bottom).

To illustrate this, we show calculations made on an interval of length 1 with periodic boundary conditions, space step  $\Delta x = 0.01$ , time step  $\Delta t = 0.9 * \Delta x$ , velocity  $V = 1$  and final time  $T = 5$ . Two types of initial conditions are tested: first a very

regular initial condition, a sine wave, then a discontinuous initial condition, a square wave (see Figure 2.5). The schemes accurate with order 1 are clearly diffusive: they destroy the solution. The Lax–Wendroff scheme which is accurate with order 2 is very good for a regular solution but oscillates for the square wave since it is dispersive. The concept of the equivalent equation allows us to understand these numerical phenomena.

**Exercise 2.3.8** Take the equation

$$\begin{cases} \frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} - \nu \frac{\partial^2 u}{\partial x^2} - \mu \frac{\partial^3 u}{\partial x^3} = 0 & \text{for } (x, t) \in \mathbb{R} \times \mathbb{R}_*^+ \\ u(t = 0, x) = \sin(\omega x + \phi) & \text{for } x \in \mathbb{R}, \end{cases}$$

with  $V, \nu, \mu, \omega, \phi \in \mathbb{R}$ . Show that its solution is

$$u(t, x) = \exp(-\nu\omega^2 t) \sin(\omega(x - (V + \mu\omega^2)t) + \phi)$$

(we shall assume uniqueness). Deduce that the diffusion attenuates the amplitude of the solution, while the dispersion modifies the velocity of propagation.

**Exercise 2.3.9** Define the ‘leapfrog’ scheme

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0.$$

Study the consistency and the truncation error of this scheme. Show by Fourier analysis that it is stable under the condition CFL  $|V|\Delta t \leq M\Delta x$  with  $M < 1$ .

**Exercise 2.3.10** Define the Crank–Nicolson scheme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{4\Delta x} + V \frac{u_{j+1}^n - u_{j-1}^n}{4\Delta x} = 0.$$

Study the consistency and the truncation error of this scheme. Show by Fourier analysis that it is unconditionally stable.

## 2.3.2 Wave equation

We consider the wave equation in the bounded domain  $(0, 1)$  with periodic boundary conditions

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0 & \text{for } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ u(t, x + 1) = u(t, x) & \text{for } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ u(t = 0, x) = u_0(x) & \text{for } x \in (0, 1) \\ \frac{\partial u}{\partial t}(t = 0, x) = u_1(x) & \text{for } x \in (0, 1). \end{cases} \quad (2.40)$$

With the same notation as above, the discrete unknown at each time step is a vector  $u^n = (u_j^n)_{0 \leq j \leq N} \in \mathbb{R}^{N+1}$ . The periodic boundary conditions lead to equations  $u_0^n = u_{N+1}^n$  for all  $n \geq 0$ , and more generally  $u_j^n = u_{N+1+j}^n$ . As the boundary conditions do not fix the value of  $u$  at the end of the interval  $(0, 1)$  (think of the interpretation in terms of a vibrating cord), the solution  $u$  cannot remain bounded in time, which complicates the study of the stability of the numerical schemes. For example, if  $u_0 \equiv 0$  and  $u_1 \equiv C$  in  $(0, 1)$ , the solution of (2.40) is  $u(t, x) = Ct$ . To eliminate this effect, we make the hypothesis that the initial velocity is on average zero

$$\int_0^1 u_1(x) dx = 0. \quad (2.41)$$

For the wave equation (2.40) the usual scheme is the  **$\theta$ -centred scheme**: for  $n \geq 1$  and  $j \in \{0, \dots, N\}$ ,

$$\begin{aligned} & \frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{(\Delta t)^2} + \theta \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} \\ & + (1 - 2\theta) \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} + \theta \frac{-u_{j-1}^{n-1} + 2u_j^{n-1} - u_{j+1}^{n-1}}{(\Delta x)^2} = 0 \end{aligned} \quad (2.42)$$

with  $0 \leq \theta \leq 1/2$ . When  $\theta = 0$  we obtain an explicit scheme, while the scheme is implicit if  $\theta \neq 0$ . The initial conditions are taken into account by

$$u_j^0 = u_0(x_j) \quad \text{and} \quad \frac{u_j^1 - u_j^0}{\Delta t} = \int_{x_{j-1/2}}^{x_{j+1/2}} u_1(x) dx,$$

which guarantees that the discrete initial velocity also satisfies the condition (2.41). As each of the centred finite differences which approximate the second derivatives in (2.42) is of order 2, the  $\theta$ -centred scheme (2.42) has accuracy of order 2 in space and time. We remark that this scheme is invariant if we change the sense of time (which is compatible with the reversibility property in time of the wave equation, see in Section 1.3.2).

**Lemma 2.3.6** *If  $1/4 \leq \theta \leq 1/2$ , the  $\theta$ -centred scheme (2.42) is unconditionally stable in the  $L^2$  norm. If  $0 \leq \theta < 1/4$ , it is stable under the CFL condition*

$$\frac{\Delta t}{\Delta x} < \sqrt{\frac{M}{1 - 4\theta}}, \quad \text{with } 0 < M < 1,$$

*and unstable if  $\Delta t/\Delta x > 1/\sqrt{1 - 4\theta}$ .*

**Proof.** As before, we use Fourier analysis to obtain

$$\hat{u}^{n+1}(k) - 2\hat{u}^n(k) + \hat{u}^{n-1}(k) + \alpha(k) (\theta \hat{u}^{n+1}(k) + (1 - 2\theta)\hat{u}^n(k) + \theta \hat{u}^{n-1}(k)) = 0,$$

with

$$\alpha(k) = 4 \left( \frac{\Delta t}{\Delta x} \right)^2 \sin^2(\pi k \Delta x).$$

This is a three level scheme which we rewrite as

$$\begin{aligned} \hat{U}^{n+1}(k) &= \begin{pmatrix} \hat{u}^{n+1}(k) \\ \hat{u}^n(k) \end{pmatrix} = \begin{pmatrix} (2 - (1 - 2\theta)\alpha(k))/1 + \theta\alpha(k) & -1 \\ 1 & 0 \end{pmatrix} \hat{U}^n(k) \\ &= A(k)\hat{U}^n(k), \end{aligned}$$

and  $\hat{U}^{n+1}(k) = A(k)^n \hat{U}^1(k)$ . The eigenvalues  $(\lambda_1, \lambda_2)$  of the matrix  $A(k)$  are the roots of the second degree polynomial

$$\lambda^2 - \frac{2 - (1 - 2\theta)\alpha(k)}{1 + \theta\alpha(k)}\lambda + 1 = 0.$$

The discriminant of this equation is

$$\Delta = -\frac{\alpha(k)(4 - (1 - 4\theta)\alpha(k))}{(1 + \theta\alpha(k))^2}.$$

The study of the stability of the scheme is very delicate as  $A(k)$  is not a normal matrix and  $\|A(k)^n\|_2 \neq \rho(A(k))^n$ , where  $\rho(A(k)) = \max(|\lambda_1|, |\lambda_2|)$  is the spectral radius of  $A(k)$ . We therefore restrict ourselves to verifying the **necessary** Von Neumann stability condition,  $\rho(A(k)) \leq 1$  (see remark 2.2.24), and we refer to exercise 2.3.11 for a sufficient condition. If  $\Delta t/\Delta x > 1/\sqrt{1 - 4\theta}$ , a judicious choice of  $k$  (such that  $\sin^2(\pi k \Delta x) \approx 1$ ) leads to  $\Delta > 0$ , and in this case the two roots  $\lambda_1$  and  $\lambda_2$  are real, with product equal to 1. One of the two must be strictly greater than 1 in modulus,  $\rho(A(k)) > 1$ , and the scheme is therefore unstable. If  $\Delta t/\Delta x < 1/\sqrt{1 - 4\theta}$  or  $\theta \geq \frac{1}{4}$ , then  $\Delta \leq 0$  for all  $k$ , and the two roots are complex conjugate with modulus equal to 1. Consequently,  $\rho(A(k)) = 1$  and the Von Neumann stability condition is satisfied.  $\square$

**Exercise 2.3.11** Finish the proof of Lemma 2.3.6 by calculating  $A(k)^n$ , and show the stability of the scheme under the CFL condition, thanks to (2.41).

**Exercise 2.3.12** We consider the limiting case of the Lemma 2.3.6, that is,  $\Delta t/\Delta x = 1/\sqrt{1 - 4\theta}$  with  $0 \leq \theta < 1/4$ . Show that the  $\theta$ -centred scheme (2.42) is unstable in this case by verifying that  $u_j^n = (-1)^{n+j}(2n - 1)$  is a solution (note that this is ‘weak’ instability since the increase of  $u^n$  is linear and nonexponential).

We illustrate these schemes in Figure 2.6 on which we show the results obtained with the explicit centred scheme and the  $\theta$ -implicit scheme (for  $\theta = 0.25$ ). The calculations are made on an interval of length 1 with periodic boundary conditions, space step  $\Delta x = 0.01$ , time step  $\Delta t = 0.9 * \Delta x$ , and final time  $T = 5$ . The initial condition  $u_0$  is a sine wave, while  $u_1$  is zero.

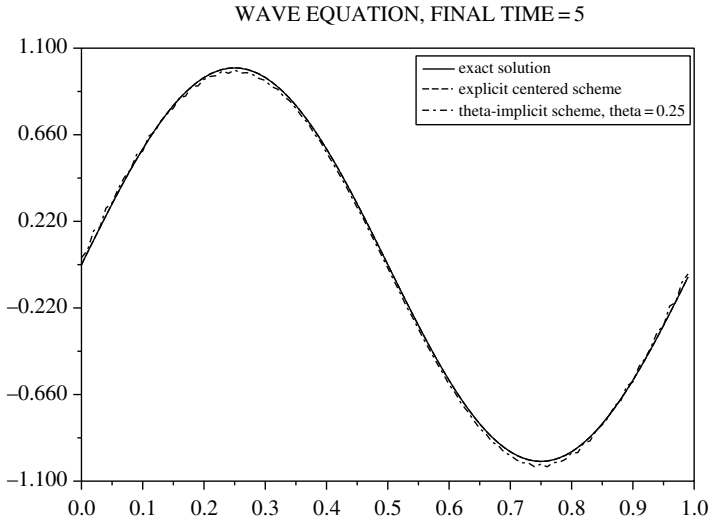


Figure 2.6. Schemes for the wave equation.

We have seen in exercise 1.3.4 that the wave equation (2.40) satisfies a conservation of energy property, that is, for all  $t > 0$ ,

$$E(t) = E(0) \quad \text{with} \quad E(t) = \int_0^1 \left| \frac{\partial u}{\partial t}(t, x) \right|^2 dx + \int_0^1 \left| \frac{\partial u}{\partial x}(t, x) \right|^2 dx.$$

It is often desirable that a numerical scheme satisfies (exactly or approximately) a discrete version of this conservation of energy. For the  $\theta$ -scheme we introduce the **discrete energy**

$$E^{n+1} = \sum_{j=0}^N \left( \frac{u_j^{n+1} - u_j^n}{\Delta t} \right)^2 + a_{\Delta x}(u^{n+1}, u^n) + \theta a_{\Delta x}(u^{n+1} - u^n, u^{n+1} - u^n)$$

with

$$a_{\Delta x}(u, v) = \sum_{j=0}^N \left( \frac{u_{j+1} - u_j}{\Delta x} \right) \left( \frac{v_{j+1} - v_j}{\Delta x} \right).$$

Clearly,  $E^{n+1}$  is an approximation, to  $\mathcal{O}(\Delta x + \Delta t)$ , of the exact energy  $E(t_{n+1})$ . We leave it the reader to prove the property of the conservation of discrete energy.

**Exercise 2.3.13** Show that the  $\theta$ -centred scheme (2.42) conserves the discrete energy, that is,  $E^n = E^1$  for all  $n \geq 0$ .

Another way to define the schemes for the wave equation is to start by rewriting (2.40) as a system of first order equations. Introducing  $v = \partial u / \partial t$  and  $w = \partial u / \partial x$ , (2.40) is equivalent to

$$\left\{ \begin{array}{ll} \frac{\partial}{\partial t} \begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} v \\ w \end{pmatrix} & \text{for } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ v(t, x+1) = v(t, x), w(t, x+1) = w(t, x) & \text{for } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ w(t=0, x) = \frac{\partial u_0}{\partial x}(x) & \text{for } x \in (0, 1) \\ v(t=0, x) = u_1(x) & \text{for } x \in (0, 1). \end{array} \right. \quad (2.43)$$

We can give a physical or mechanical interpretation of these new variables. If  $u$  models a displacement (of the vibrating cord, for example), then  $v$  is a velocity and  $w$  is a deformation. The system of two equations (2.43) occurs as a generalization of the advection equation. We can therefore define a **Lax–Friedrichs**-type scheme

$$\frac{1}{2\Delta t} \begin{pmatrix} 2v_j^{n+1} - v_{j+1}^n - v_{j-1}^n \\ 2w_j^{n+1} - w_{j+1}^n - w_{j-1}^n \end{pmatrix} - \frac{1}{2\Delta x} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v_{j+1}^n - v_{j-1}^n \\ w_{j+1}^n - w_{j-1}^n \end{pmatrix} = 0, \quad (2.44)$$

or one of **Lax–Wendroff** type

$$\begin{aligned} & \frac{1}{\Delta t} \begin{pmatrix} v_j^{n+1} - v_j^n \\ w_j^{n+1} - w_j^n \end{pmatrix} - \frac{1}{2\Delta x} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v_{j+1}^n - v_{j-1}^n \\ w_{j+1}^n - w_{j-1}^n \end{pmatrix} \\ & + \frac{\Delta t}{2(\Delta x)^2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^2 \begin{pmatrix} -v_{j-1}^n + 2v_j^n - v_{j+1}^n \\ -w_{j-1}^n + 2w_j^n - w_{j+1}^n \end{pmatrix} = 0. \end{aligned} \quad (2.45)$$

**Exercise 2.3.14** Show that the Lax–Friedrichs scheme (2.44) is stable in the  $L^2$  norm under the CFL condition  $\Delta t \leq \Delta x$ , and that it is accurate with order 1 in space and time if the ratio  $\Delta t / \Delta x$  is held constant as  $\Delta t$  and  $\Delta x$  tend to zero.

**Exercise 2.3.15** Show that the Lax–Wendroff scheme (2.45) is stable in the  $L^2$  norm under the CFL condition  $\Delta t \leq \Delta x$ , and that it is accurate with order 2 in space and time.

As for the advection equation, a fundamental idea to obtain ‘good’ schemes is **upwinding**. However, here we have a system of two equations and it is not clear which is the velocity, which is what allows us to upwind. In fact, it is sufficient to diagonalize the matrix

$$J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

to obtain two decoupled advection equations which are upwinded in different ways from each other. It is therefore the eigenvalues of the matrix  $J$  (1 and  $-1$ , in fact) which play the role of the velocity, and we upwind component by component in the decomposition in a basis of eigenvectors of this matrix. This type of scheme is used for hyperbolic systems and, in particular, for gas dynamics (see [23] to which we refer for more detail).



*This page intentionally left blank*

# 3 Variational formulation of elliptic problems

---

## 3.1 Generalities

### 3.1.1 Introduction

In this chapter we are interested in the mathematical analysis of **elliptic partial differential equations** (PDEs) (see definition 1.5.5). In general, these elliptic equations correspond to stationary physical models, that is, models which are independent of time. We shall see that boundary value problems are well-posed for these elliptic PDEs, that is, they have a solution which is unique and depends continuously on the data. The approach that we shall follow is called the **variational approach**. First we should say that the interest of this approach goes far beyond the framework of elliptic PDEs and even the framework of the ‘pure’ mathematical analysis to which we restrict ourselves. Indeed, we shall return to this variational approach for problems of evolution in time (parabolic or hyperbolic PDEs), and it will be crucial for understanding the finite element method that we develop in Chapter 6. Additionally, this approach has a very natural physical or mechanical interpretation. The reader should make the effort to study this variational approach carefully!

In this chapter and the following, the prototype example of elliptic PDEs will be the Laplacian for which we shall study the following boundary value problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (3.1)$$

where we impose Dirichlet boundary conditions (we refer to Section 1.3.3 for a presentation of this model). In (3.1),  $\Omega$  is an open set of the space  $\mathbb{R}^N$ ,  $\partial\Omega$  is its

boundary,  $f$  is a right-hand side data for the problem, and  $u$  is the unknown. Of course, in Chapter 5 we shall give many other examples of elliptic PDEs which can be studied, thanks to the variational approach.

The plan of this chapter is the following. In Section 3.2 we recall some integration by parts formulas, called **Green's formulas**, then we define the **variational formulation**. Section 3.3 is dedicated to **Lax–Milgram theorem** which will be the essential tool allowing us to show existence and uniqueness of the solutions of the variational formulation. We shall see that, to apply this theorem, it is inescapable that we must give up the space  $C^1(\overline{\Omega})$  of continuously differentiable functions and use its ‘generalization’, the Sobolev space  $H^1(\Omega)$ .

We conclude this introduction by mentioning other methods to solve PDEs which are less powerful or more complicated than the variational approach (we refer the curious, and courageous, reader to the encyclopaedia [14]).

### 3.1.2 Classical formulation

The ‘classical’ formulation of (3.1), which might appear ‘natural’ at first sight, is to assume sufficient regularity for the solution  $u$  so that equations (3.1) have a meaning at every point of  $\Omega$  or of  $\partial\Omega$ . First we recall some notation related to spaces of regular functions.

**Definition 3.1.1** *Let  $\Omega$  be an open set of  $\mathbb{R}^N$ , and  $\overline{\Omega}$  its closure. We denote by  $C(\Omega)$  (respectively,  $C(\overline{\Omega})$ ) the space of continuous function in  $\Omega$  (respectively, in  $\overline{\Omega}$ ). Let  $k \geq 0$  be an integer. We denote by  $C^k(\Omega)$  (respectively,  $C^k(\overline{\Omega})$ ) the space of functions  $k$  times continuously differentiable in  $\Omega$  (respectively, in  $\overline{\Omega}$ ).*

A **classical solution** (we also say **strong solution**) of (3.1) is a solution  $u \in C^2(\Omega) \cap C(\overline{\Omega})$ , which implies that the right-hand side  $f$  must be in  $C(\Omega)$ . This classical formulation, unfortunately, has a number of problems! Without going into detail, we note that, under the single hypothesis  $f \in C(\overline{\Omega})$ , there is not in general a solution of class  $C^2$  for (3.1) if the dimension of the space is greater than two ( $N \geq 2$ ). In fact, a solution does exist, as we shall see later, but it is not of class  $C^2$  (it is a little less regular except if the data  $f$  is more regular than  $C(\overline{\Omega})$ ). The case of a space with dimension one ( $N = 1$ ) is particular as it is easy to find classical solutions (see exercise 3.1.1), but we shall, nevertheless, see that, even in this successful case, the classical formulation is inconvenient.

In what follows, to study (3.1), we shall replace its classical formulation by a so-called variational formulation, which is much more advantageous.

### 3.1.3 The case of a space of one dimension

In one space dimension ( $N = 1$ ), if  $\Omega = (0, 1)$ , the boundary value problem (3.1) becomes

$$\begin{cases} -\frac{d^2u}{dx^2} = f & \text{for } 0 < x < 1 \\ u(0) = u(1) = 0. \end{cases} \quad (3.2)$$

This problem is so simple it has an explicit solution!

**Exercise 3.1.1** If  $f$  is a continuous function in  $[0, 1]$ , show that (3.2) has a unique solution in  $C^2([0, 1])$  given by the formula

$$u(x) = x \int_0^1 f(s)(1-s)ds - \int_0^x f(s)(x-s)ds \quad \text{for } x \in [0, 1]. \quad (3.3)$$

For the remainder of this section we shall forget the explicit formula (3.3) which does not have any equivalent for more complicated problems.

In one space dimension, ‘partial differential equation’ loses its meaning as, since we only have one variable, we can more simply say ‘ordinary differential equation’. However, equation (3.2) is not a ‘normal’ ordinary differential equation in the sense that the solution must satisfy conditions ‘at both ends’ rather than an initial condition at one end of the interval  $[0, 1]$ . This is exactly the difference between a boundary value problem (with conditions ‘at both ends’) and a Cauchy problem (with an initial condition ‘at one end’).

It is, however, interesting to see that, even in one dimension, the classical methods of ordinary differential equations are not useful to study (3.2) (and are completely useless in higher dimensions). For a parameter  $m \in \mathbb{R}$ , we consider the Cauchy problem for the Laplacian with initial data at 0

$$\begin{cases} -\frac{d^2u}{dx^2} = f & \text{for } 0 < x < 1 \\ u(0) = 0, \quad \frac{du}{dx}(0) = m. \end{cases} \quad (3.4)$$

Obviously there is a unique solution of (3.4): it is enough to integrate this linear equation (or more generally to use the Cauchy–Lipschitz existence theorem). It is not at all clear, on the other hand, that the solution of (3.4) coincides with that of (3.2) (if it exists). We ask the question if there exists a parameter  $m$  such that the solution of (3.4) also satisfies  $u(1) = 0$  and therefore is a solution of (3.2). This is the principle of the **shooting method** which allows us to solve, both theoretically and numerically, the boundary value problem (3.2). Iteratively, we predict a value of  $m$  (shooting from the point 0), we integrate the Cauchy problem (3.4) (we calculate the trajectory of the shot), then depending on the result  $u(1)$  we correct the value of  $m$ . In practice, this is not a very effective method which has the major problem that it cannot be generalized to higher dimensions.

The conclusion is that we need methods specific to boundary value problems which have nothing to do with those related to Cauchy problems.

## 3.2 Variational approach

The principle of the variational approach for the solution of PDEs is to replace the equation by an equivalent so-called variational formulation obtained by integrating the equation multiplied by an arbitrary function, called a test function. As we need to carry out integration by parts when establishing the variational formulation, we start by giving some essential results on this subject.

### 3.2.1 Green's formulas

In this section  $\Omega$  is an open set of the space  $\mathbb{R}^N$  (which may be bounded or unbounded), whose boundary is denoted by  $\partial\Omega$ . We also assume that  $\Omega$  is a **regular** open set of class  $\mathcal{C}^1$ . The precise definition of a regular open set is given below in definition 3.2.5, but it is not necessary to understand this absolutely to follow the rest of this course. It is enough to know that an open regular set is *roughly speaking* an open set whose boundary is a regular hypersurface (a manifold of dimension  $N - 1$ ), and this open set is locally situated on one side of its boundary. We then define the **outward normal** at the boundary  $\partial\Omega$  as being the unit vector  $n = (n_i)_{1 \leq i \leq N}$  normal at every point to the tangent plane of  $\Omega$  and pointing to the exterior of  $\Omega$  (see Figure 1.1). In  $\Omega \subset \mathbb{R}^N$  we denote by  $dx$  the volume measure, or Lebesgue measure of dimension  $N$ . On  $\partial\Omega$ , we denote by  $ds$  the surface measure, or Lebesgue measure of dimension  $N - 1$  on the manifold  $\partial\Omega$ . The principal result of this section is the following theorem (see [4], [38]).

**Theorem 3.2.1 (Green's formula)** *Let  $\Omega$  be a regular open set of class  $\mathcal{C}^1$ . Let  $w$  be a  $C^1(\overline{\Omega})$  function with bounded support in the closure  $\overline{\Omega}$ . Then  $w$  satisfies Green's formula*

$$\int_{\Omega} \frac{\partial w}{\partial x_i}(x) dx = \int_{\partial\Omega} w(x) n_i(x) ds, \quad (3.5)$$

where  $n_i$  is the  $i$ th component of the unit outward normal to  $\Omega$ .

**Remark 3.2.2** To say that a regular function  $w$  has bounded support in the closed set  $\overline{\Omega}$  is the same as saying that it is zero at infinity if the closed set is unbounded. We also say that the function  $w$  has compact support in  $\overline{\Omega}$  (take care: this does not imply that  $w$  is zero on the boundary  $\partial\Omega$ ). In particular, the hypothesis of theorem 3.2.1 in connection with the bounded support of the function  $w$  in  $\overline{\Omega}$  is pointless if the open set  $\Omega$  is bounded. If  $\Omega$  is unbounded, this hypothesis ensures that the integrals in (3.5) are finite •

Theorem 3.2.1 has many corollaries which are all immediate consequences of Green's formula (3.5). The reader who wants to save his memory need only remember Green's formula (3.5)!

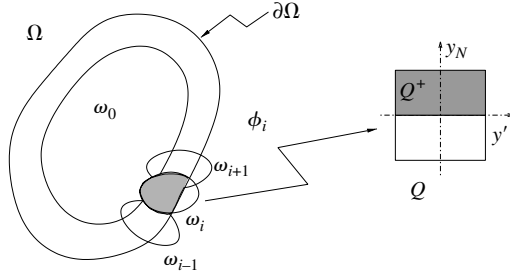


Figure 3.1. Definition of the regularity of an open set.

**Corollary 3.2.3 (Integration by parts formula)** *Let  $\Omega$  be a regular open set of class  $C^1$ . Let  $u$  and  $v$  be two  $C^1(\overline{\Omega})$  functions with bounded support in the closed set  $\overline{\Omega}$ . Then they satisfy the integration by parts formula*

$$\int_{\Omega} u(x) \frac{\partial v}{\partial x_i}(x) dx = - \int_{\Omega} v(x) \frac{\partial u}{\partial x_i}(x) dx + \int_{\partial\Omega} u(x) v(x) n_i(x) ds. \quad (3.6)$$

**Proof.** It is enough to take  $w = uv$  in theorem 3.2.1.  $\square$

**Corollary 3.2.4** *Let  $\Omega$  be a regular open set of class  $C^1$ . Let  $u$  be a function of  $C^2(\overline{\Omega})$  and  $v$  a function of  $C^1(\overline{\Omega})$ , both with bounded support in the closed set  $\overline{\Omega}$ . Then they satisfy the integration by parts formula*

$$\int_{\Omega} \Delta u(x) v(x) dx = - \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \frac{\partial u}{\partial n}(x) v(x) ds, \quad (3.7)$$

where  $\nabla u = \left( \frac{\partial u}{\partial x_i} \right)_{1 \leq i \leq N}$  is the gradient vector of  $u$ , and  $\frac{\partial u}{\partial n} = \nabla u \cdot n$ .

**Proof.** We apply corollary 3.2.3 to  $v$  and  $\frac{\partial u}{\partial x_i}$  and we sum in  $i$ .  $\square$

**Definition 3.2.5** *We say that an open set  $\Omega$  of  $\mathbb{R}^N$  is regular of class  $C^k$  (for an integer  $k \geq 1$ ) if there exist a finite number of open sets  $(\omega_i)_{0 \leq i \leq I}$  such that*

$$\overline{\omega_0} \subset \Omega, \quad \overline{\Omega} \subset \cup_{i=0}^I \omega_i, \quad \partial\Omega \subset \cup_{i=1}^I \omega_i,$$

and that, for every  $i \in \{1, \dots, I\}$  (see Figure 3.1), there exists a bijective mapping  $\phi_i$  of class  $C^k$ , from  $\omega_i$  into the set

$$Q = \{y = (y', y_N) \in \mathbb{R}^{N-1} \times \mathbb{R}, |y'| < 1, |y_N| < 1\},$$

whose inverse is also of class  $C^k$ , and such that

$$\begin{aligned}\phi_i(\omega_i \cap \Omega) &= Q \cap \{y = (y', y_N) \in \mathbb{R}^{N-1} \times \mathbb{R}, y_N > 0\} = Q^+, \\ \phi_i(\omega_i \cap \partial\Omega) &= Q \cap \{y = (y', y_N) \in \mathbb{R}^{N-1} \times \mathbb{R}, y_N = 0\}.\end{aligned}$$

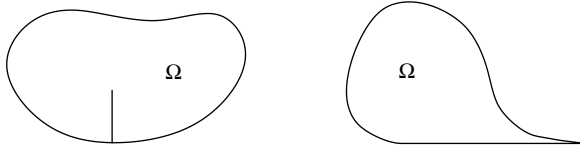


Figure 3.2. Two examples of a nonregular open set: open set with a crack on the left, open set with a cusp on the right.

**Remark 3.2.6** Even though Figure 3.1 represents a bounded regular open set, the definition 3.2.5 also applies to unbounded open sets. Definition 3.2.5 does not only exclude open sets whose boundary is not a regular surface, but it also excludes open sets which do not lie locally on one side of their boundary. Figure 3.2 contains two typical examples of a nonregular open set which give an irremovable singularity, a crack, and a cusp. These examples are not ‘mathematical inventions’: the cracked set is used to study crack problems in structural mechanics. We can, nevertheless, generalize the class of regular open set a little to open sets which are ‘piecewise regular’, provided that the pieces of the boundary are ‘joined’ by angles different from either 0 (a cusp) or from  $2\pi$  (a crack). All of these details are largely outside the scope of this course, and we refer the reader to remark 4.3.7 for another explanation of regularity problems. •

**Exercise 3.2.1** From Green’s formula (3.5) deduce the Stokes formula

$$\int_{\Omega} \operatorname{div} \sigma(x) \phi(x) dx = - \int_{\Omega} \sigma(x) \cdot \nabla \phi(x) dx + \int_{\partial\Omega} \sigma(x) \cdot n(x) \phi(x) ds,$$

where  $\phi$  is a scalar function of  $C^1(\overline{\Omega})$  and  $\sigma$  a vector valued function of  $C^1(\overline{\Omega})$ , with bounded supports in the closed set  $\overline{\Omega}$ .

**Exercise 3.2.2** In  $N = 3$  dimensions we define the curl of a function of  $\Omega$  in  $\mathbb{R}^3$ ,  $\phi = (\phi_1, \phi_2, \phi_3)$ , as the function of  $\Omega$  in  $\mathbb{R}^3$  defined by

$$\nabla \times \phi = \left( \frac{\partial \phi_3}{\partial x_2} - \frac{\partial \phi_2}{\partial x_3}, \frac{\partial \phi_1}{\partial x_3} - \frac{\partial \phi_3}{\partial x_1}, \frac{\partial \phi_2}{\partial x_1} - \frac{\partial \phi_1}{\partial x_2} \right).$$

For  $\phi$  and  $\psi$ , vector valued functions of  $C^1(\overline{\Omega})$ , with bounded supports in the closed set  $\overline{\Omega}$ , deduce Green’s formula (3.5)

$$\int_{\Omega} \nabla \times \phi \cdot \psi dx - \int_{\Omega} \phi \cdot \nabla \times \psi dx = - \int_{\partial\Omega} (\phi \times n) \cdot \psi ds.$$

### 3.2.2 Variational formulation

To simplify the presentation, we assume that the open set  $\Omega$  is bounded and regular, and that the right-hand side  $f$  of (3.1) is continuous on  $\overline{\Omega}$ . The principal result of this section is the following proposition.

**Proposition 3.2.7** *Let  $u$  be a function of  $C^2(\overline{\Omega})$ . Let  $X$  be the space defined by*

$$X = \{\phi \in C^1(\overline{\Omega}) \text{ such that } \phi = 0 \text{ on } \partial\Omega\}.$$

*Then  $u$  is a solution of the boundary value problem (3.1) if and only if  $u$  belongs to  $X$  and satisfies the equation*

$$\int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx = \int_{\Omega} f(x)v(x) dx \quad \text{for every } v \in X. \quad (3.8)$$

*Equation (3.8) is called the **variational formulation** of the boundary value problem (3.1).*

**Remark 3.2.8** An immediate consequence of the variational formulation (3.8) is that it is meaningful if the solution  $u$  is only a function of  $C^1(\overline{\Omega})$ , as opposed to the ‘classical’ formulation (3.1) which requires  $u$  to belong to  $C^2(\overline{\Omega})$ . We therefore already suspect that it is easier to solve (3.8) than (3.1) since it is less demanding on the regularity of the solution.

In the variational formulation (3.8), the function  $v$  is called the **test function**. The variational formulation is also sometimes called the weak form of the boundary value problem (3.1). In mechanics, the variational formulation is known as the ‘principle of virtual work’. In physics, we also talk of the balance equation or the reciprocity formula.

When we take  $v = u$  in (3.8), we obtain what is called an **energy equality**, which in general expresses the equality between the stored energy in the domain  $\Omega$  (the left-hand term of (3.8)) and a potential energy associated with  $f$  (the right-hand term of (3.8)). •

**Proof.** If  $u$  is a solution of the boundary value problem (3.1), we multiply the equation by  $v \in X$  and we use the integration by parts formula of corollary 3.2.4.

$$\int_{\Omega} \Delta u(x)v(x) dx = - \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \frac{\partial u}{\partial n}(x)v(x) ds,$$

where  $v = 0$  on  $\partial\Omega$  since  $v \in X$ , therefore

$$\int_{\Omega} f(x)v(x) dx = \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx,$$

which is nothing other than the formula (3.8). Conversely, if  $u \in X$  satisfies (3.8), by using the integration by parts formula ‘in reverse’ we obtain

$$\int_{\Omega} (\Delta u(x) + f(x))v(x) dx = 0 \quad \text{for every } v \in X.$$



As  $(\Delta u + f)$  is a continuous function, thanks to lemma 3.2.9 we conclude that  $-\Delta u(x) = f(x)$  for all  $x \in \Omega$ . In addition, since  $u \in X$ , we recover the boundary condition  $u = 0$  on  $\partial\Omega$ , that is,  $u$  is a solution of the boundary value problem (3.1).  $\square$

**Lemma 3.2.9** *Let  $\Omega$  be an open set of  $\mathbb{R}^N$ . Let  $g(x)$  be a continuous function in  $\Omega$ . If for every function  $\phi$  of  $C^\infty(\Omega)$  with compact support in  $\Omega$ , we have*

$$\int_{\Omega} g(x)\phi(x) dx = 0,$$

*then the function  $g$  is zero in  $\Omega$ .*

**Proof.** Assume that there exists a point  $x_0 \in \Omega$  such that  $g(x_0) \neq 0$ . Without loss of generality, we can assume that  $g(x_0) > 0$  (otherwise we take  $-g$ ). By continuity, there exists a small open neighbourhood  $\omega \subset \Omega$  of  $x_0$  such that  $g(x) > 0$  for all  $x \in \omega$ . Let  $\phi$  be a nonzero positive test function with support in  $\omega$ . We have

$$\int_{\Omega} g(x)\phi(x) dx = \int_{\omega} g(x)\phi(x) dx = 0,$$

which contradicts the hypothesis on  $g$ . Therefore  $g(x) = 0$  for all  $x \in \Omega$ .  $\square$

**Remark 3.2.10** We can rewrite the variational formulation (3.8) in compact notation: find  $u \in X$  such that

$$a(u, v) = L(v) \quad \text{for every } v \in X,$$

with

$$a(u, v) = \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx$$

and

$$L(v) = \int_{\Omega} f(x)v(x) dx,$$

where  $a(\cdot, \cdot)$  is a bilinear form on  $X$  and  $L(\cdot)$  is a linear form on  $X$ . It is in this abstract form that we solve (with some hypotheses) the variational formulation in the next section.  $\bullet$

The principle idea of **the variational approach** is to show the existence and uniqueness of the solution of the variational formulation (3.8), which implies the same result for the equation (3.1) because of proposition 3.2.7. Indeed, we shall see that there is a theory, both simple and powerful, for analysing variational formulations. Nonetheless, this theory only works if the space in which we look for the solution and in which we take the test functions (in the preceding notation, the space  $X$ ) is a Hilbert space, which is not the case for  $X = \{v \in C^1(\overline{\Omega}), v = 0 \text{ on } \partial\Omega\}$  equipped with the ‘natural’ scalar product for this problem. The main difficulty in the application of the variational approach will therefore be that we must use a space other than  $X$ , that is the Sobolev space  $H_0^1(\Omega)$  which is indeed a Hilbert space (see Chapter 4).

**Exercise 3.2.3** In a bounded open set  $\Omega$  we consider the Laplacian with Neumann boundary condition

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases} \quad (3.9)$$

Let  $u$  be a function of  $C^2(\overline{\Omega})$ . Show that  $u$  is a solution of the boundary value problem (3.9) if and only if  $u$  satisfies the equation

$$\int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx = \int_{\Omega} f(x)v(x) dx \quad \text{for every } v \in C^1(\overline{\Omega}). \quad (3.10)$$

Deduce from this that a necessary condition for the existence of a solution in  $C^2(\overline{\Omega})$  of (3.9) is that  $\int_{\Omega} f(x)dx = 0$ .

**Exercise 3.2.4** In a bounded open set  $\Omega$  we consider the plate equation

$$\begin{cases} \Delta(\Delta u) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega \end{cases} \quad (3.11)$$

We denote by  $X$  the space of functions  $v$  of  $C^2(\overline{\Omega})$  such that  $v$  and  $\frac{\partial v}{\partial n}$  are zero on  $\partial\Omega$ . Let  $u$  be a function of  $C^4(\overline{\Omega})$ . Show that  $u$  is a solution of the boundary value problem (3.11) if and only if  $u$  belongs to  $X$  and satisfies the equation

$$\int_{\Omega} \Delta u(x) \Delta v(x) dx = \int_{\Omega} f(x)v(x) dx \quad \text{for every } v \in X. \quad (3.12)$$

## 3.3 Lax–Milgram theory

### 3.3.1 Abstract framework

We describe an abstract theory to obtain the existence and the uniqueness of the solution of a variational formulation in a Hilbert space. We denote by  $V$  a real Hilbert space with scalar product  $\langle \cdot, \cdot \rangle$  and norm  $\| \cdot \|$ . Following remark 3.2.10 we consider a variational formulation of the type:

$$\text{find } u \in V \text{ such that } a(u, v) = L(v) \text{ for every } v \in V. \quad (3.13)$$

The hypotheses on  $a$  and  $L$  are

- (1)  $L(\cdot)$  is a continuous linear form on  $V$ , that is,  $v \rightarrow L(v)$  is linear from  $V$  into  $\mathbb{R}$  and there exists  $C > 0$  such that

$$|L(v)| \leq C\|v\| \quad \text{for all } v \in V;$$

- (2)  $a(\cdot, \cdot)$  is a bilinear form on  $V$ , that is,  $w \rightarrow a(w, v)$  is a linear form from  $V$  into  $\mathbb{R}$  for all  $v \in V$ ; and  $v \rightarrow a(w, v)$  is a linear form from  $V$  into  $\mathbb{R}$  for all  $w \in V$ ;

(3)  $a(\cdot, \cdot)$  is continuous, that is, there exists  $M > 0$  such that

$$|a(w, v)| \leq M \|w\| \|v\| \quad \text{for all } w, v \in V; \quad (3.14)$$

(4)  $a(\cdot, \cdot)$  is **coercive** (or elliptic), that is, there exists  $\nu > 0$  such that

$$a(v, v) \geq \nu \|v\|^2 \quad \text{for all } v \in V. \quad (3.15)$$

As we shall see in this section, all the hypotheses above are necessary to solve (3.13). In particular, the coercivity of  $a(\cdot, \cdot)$  is essential.

**Theorem 3.3.1 (Lax–Milgram)** *Let  $V$  be a real Hilbert space,  $L(\cdot)$  a continuous linear form on  $V$ ,  $a(\cdot, \cdot)$  a continuous coercive bilinear form on  $V$ . Then the variational formulation (3.13) has a unique solution. Further, this solution depends continuously on the linear form  $L$ .*

**Proof.** For all  $w \in V$ , the mapping  $v \rightarrow a(w, v)$  is a continuous linear form on  $V$ : consequently, the Riesz representation theorem 12.1.18 implies that there exists an element of  $V$ , denoted  $A(w)$ , such that

$$a(w, v) = \langle A(w), v \rangle \quad \text{for all } v \in V.$$

Moreover, the bilinearity of  $a(w, v)$  obviously implies the linearity of the mapping  $w \rightarrow A(w)$ . Further, by taking  $v = A(w)$ , the continuity (3.14) of  $a(w, v)$  shows that

$$\|A(w)\|^2 = a(w, A(w)) \leq M \|w\| \|A(w)\|,$$

that is,  $\|A(w)\| \leq M \|w\|$  and therefore  $w \rightarrow A(w)$  is continuous. Another application of the Riesz representation theorem 12.1.18 implies that there exists an element of  $V$ , denoted  $f$ , such that  $\|f\|_V = \|L\|_{V'}$  and

$$L(v) = \langle f, v \rangle \quad \text{for all } v \in V.$$

Finally, the variational problem (3.13) is equivalent to:

$$\text{find } u \in V \text{ such that } A(u) = f. \quad (3.16)$$

To prove the theorem we must therefore show that the operator  $A$  is bijective from  $V$  to  $V$  (which implies the existence and the uniqueness of  $u$ ) and that its inverse is continuous (which proves the continuous dependence of  $u$  with respect to  $L$ ).

The coercivity (3.15) of  $a(w, v)$  shows that

$$\nu \|w\|^2 \leq a(w, w) = \langle A(w), w \rangle \leq \|A(w)\| \|w\|,$$

which gives

$$\nu \|w\| \leq \|A(w)\| \quad \text{for all } w \in V, \quad (3.17)$$

that is,  $A$  is injective. To show that  $A$  is surjective, that is,  $\text{Im}(A) = V$  (which is not obvious if  $V$  is infinite dimensional), it is enough to show that  $\text{Im}(A)$  is closed in  $V$  and that  $\text{Im}(A)^\perp = \{0\}$ . Indeed, in this case we see that  $V = \{0\}^\perp = (\text{Im}(A)^\perp)^\perp = \text{Im}(A) = \text{Im}(A)$ , which proves that  $A$  is surjective. Let  $A(w_n)$  be a sequence in  $\text{Im}(A)$  which converges to  $b$  in  $V$ . By virtue of (3.17) we have

$$\nu \|w_n - w_p\| \leq \|A(w_n) - A(w_p)\|$$

which tends to zero as  $n$  and  $p$  tend to infinity. Therefore  $w_n$  is a Cauchy sequence in the Hilbert space  $V$ , that is, it converges to a limit  $w \in V$ . Then, by continuity of  $A$  we deduce that  $A(w_n)$  converges to  $A(w) = b$ , that is,  $b \in \text{Im}(A)$  and  $\text{Im}(A)$  is therefore closed. On the other hand, let  $v \in \text{Im}(A)^\perp$ ; the coercivity (3.15) of  $a(w, v)$  implies that

$$\nu \|v\|^2 \leq a(v, v) = \langle A(v), v \rangle = 0,$$

that is,  $v = 0$  and  $\text{Im}(A)^\perp = \{0\}$ , which proves that  $A$  is bijective. Let  $A^{-1}$  be its inverse: the inequality (3.17) with  $w = A^{-1}(v)$  proves  $A^{-1}$  is continuous, therefore the solution  $u$  depends continuously on  $f$ .  $\square$

**Remark 3.3.2** If the Hilbert space  $V$  is finite dimensional (which is however never the case for the applications we shall see), the proof of the Lax–Milgram theorem 3.3.1 simplifies considerably. Indeed, in finite dimensions all linear mappings are continuous and the injectivity (3.17) of  $A$  is equivalent to its invertibility. We see, in this case, (as in the general case) that the coercivity hypothesis on the bilinear form  $a(w, v)$  is indispensable since it is this that gives the injectivity of  $A$ . Finally we remark that, if  $V = \mathbb{R}^N$ , a variational formulation is only the statement,  $\langle Au, v \rangle = \langle f, v \rangle$  for all  $v \in \mathbb{R}^N$ , of a simple linear system  $Au = f$ .  $\bullet$

**Remark 3.3.3** Another proof (a little less technical but which disguises some of the essential arguments) of the Lax–Milgram theorem 3.3.1 is the following. We begin as before until we reach the formulation (3.16) of the problem. To show the existence and uniqueness of the solution  $u$  of (3.16), we introduce an affine mapping  $T$  from  $V$  into  $V$ , defined by

$$T(w) = w - \mu(A(w) - f) \quad \text{with } \mu = \frac{\nu}{M^2},$$

which we shall show is a strict contraction, which proves the existence and the uniqueness of  $u \in V$  such that  $T(u) = u$  (from which we have the result). Indeed, we have

$$\begin{aligned} \|T(v) - T(w)\|^2 &= \|v - w - \mu A(v - w)\|^2 \\ &= \|v - w\|^2 - 2\mu \langle A(v - w), v - w \rangle + \mu^2 \|A(v - w)\|^2 \\ &= \|v - w\|^2 - 2\mu a(v - w, v - w) + \mu^2 \|A(v - w)\|^2 \\ &\leq (1 - 2\mu\nu + \mu^2 M^2) \|v - w\|^2 \\ &\leq (1 - \nu^2/M^2) \|v - w\|^2. \end{aligned}$$

$\bullet$

A variational formulation often has a physical interpretation, in particular if the bilinear form is symmetric. Indeed in this case, the solution of the variational formulation (3.13) attains the **minimum of an energy** (very natural in physics or mechanics).

**Proposition 3.3.4** *We take the hypotheses of the Lax–Milgram theorem 3.3.1. We further assume that the bilinear form is symmetric  $a(w, v) = a(v, w)$  for all  $v, w \in V$ . Let  $J(v)$  be the energy defined for  $v \in V$  be*

$$J(v) = \frac{1}{2}a(v, v) - L(v). \quad (3.18)$$

*Let  $u \in V$  be the unique solution of the variational formulation (3.13). Then  $u$  is also the unique point of the minimum of energy, that is,*

$$J(u) = \min_{v \in V} J(v).$$

*Conversely, if  $u \in V$  is a point giving an energy minimum  $J(v)$ , then  $u$  is the unique solution of the variational formulation (3.13).*

**Proof.** If  $u$  is the solution of the variational formulation (3.13), we can write (thanks to the symmetry of  $a$ )

$$J(u + v) = J(u) + \frac{1}{2}a(v, v) + a(u, v) - L(v) = J(u) + \frac{1}{2}a(v, v) \geq J(u).$$

As  $u + v$  is arbitrary in  $V$ ,  $u$  minimizes the energy  $J$  in  $V$ . Conversely, let  $u \in V$  be such that

$$J(u) = \min_{v \in V} J(v).$$

For  $v \in V$  we define a function  $j(t) = J(u + tv)$  from  $\mathbb{R}$  into  $\mathbb{R}$  (it is just a polynomial of second degree in  $t$ ). Since  $t = 0$  is a minimum of  $j$ , we deduce that  $j'(0) = 0$  which, by a simple calculation, is exactly the variational formulation (3.13).  $\square$

**Remark 3.3.5** We see later in Chapter 9 that, when the bilinear form  $a$  is symmetric, there is an argument other than the Lax–Milgram 3.3.1 theorem to prove the existence and the uniqueness of a solution of (3.13). Indeed, we shall demonstrate directly the existence of a unique minimum of the energy  $J(v)$ . By virtue of proposition 3.3.4, this shows the existence and the uniqueness of the solution of the variational formulation.  $\bullet$

### 3.3.2 Application to the Laplacian

We now try to apply the Lax–Milgram theorem 3.3.1 to the variational formulation (3.8) of the Laplacian with Dirichlet boundary conditions. This is written in the form (3.13) with

$$a(u, v) = \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx$$

and

$$L(v) = \int_{\Omega} f(x)v(x) dx,$$

where clearly  $a(\cdot, \cdot)$  is a bilinear form, and  $L(\cdot)$  a linear form. The space  $V$  (called before as  $X$ ) is

$$V = \{v \in C^1(\overline{\Omega}), v = 0 \text{ on } \partial\Omega\}. \quad (3.19)$$

As a scalar product on  $V$  we shall choose

$$\langle w, v \rangle = \int_{\Omega} \nabla w(x) \cdot \nabla v(x) dx, \quad (3.20)$$

which has for associated norm

$$\|v\| = \left( \int_{\Omega} |\nabla v(x)|^2 dx \right)^{1/2}.$$

We verify easily that (3.20) defines a scalar product on  $V$ : the only point which slows us is the property  $\|v\| = 0 \Rightarrow v = 0$ . Indeed, from the equality

$$\int_{\Omega} |\nabla v(x)|^2 dx = 0$$

we deduce that  $v$  is a constant on  $\Omega$ , and as  $v = 0$  on  $\partial\Omega$  we have  $v = 0$ . The motivation of the choice of (3.20) as scalar product is above all the fact that the bilinear form  $a(\cdot, \cdot)$  is **automatically coercive** for (3.20). In addition, we can easily check that  $a$  is continuous. To show that  $L$  is continuous, we must rely on the Poincaré inequality of lemma 3.3.6: we then have

$$\left| \int_{\Omega} f(x)v(x) dx \right| \leq \left( \int_{\Omega} |f(x)|^2 dx \right)^{1/2} \left( \int_{\Omega} |v(x)|^2 dx \right)^{1/2} \leq C\|v\|,$$

where  $C$  is a constant which depends on  $f$  but not on  $v$ . Therefore,  $L$  is continuous over  $V$ . All the of the hypotheses of the Lax–Milgram theorem 3.3.1 seem satisfied, however, we have missed one which prevents its application: the space  $V$  is not a Hilbert space since it is not complete for the norm induced by (3.20)! This obstruction does not come so much from the choice of the scalar product as the  $C^1$  regularity requirement on functions of the space  $V$ . An immediate way, which can be clarified, to solve the difficulty is to replace  $V$  by  $\overline{V}$ , its closure for the scalar product (3.20). Obviously, we have only moved the problem: what is the space  $\overline{V}$ ? The answer will be given in Chapter 4:  $\overline{V}$  is the Sobolev space  $H_0^1(\Omega)$  whose elements are no longer regular functions but only measurable. Another difficulty will be to see in what sense proposition 3.2.7 (which expresses the equivalence between the boundary value problem (3.1) and its variational formulation (3.8)) remains true when we replace the space  $V$  by  $\overline{V}$ .

We hope that we have therefore convinced the reader of the **natural and inescapable character of Sobolev spaces in the solution of variational formulations** of elliptic PDEs. We finish this chapter with a technical lemma, called the Poincaré inequality, which we used above.

**Lemma 3.3.6** *Let  $\Omega$  be an open set of  $\mathbb{R}^N$  bounded in at least one space direction. There exists a constant  $C > 0$  such that, for every function  $v \in C^1(\overline{\Omega})$  which is zero on the boundary  $\partial\Omega$ ,*

$$\int_{\Omega} |v(x)|^2 dx \leq C \int_{\Omega} |\nabla v(x)|^2 dx.$$

**Proof.** The hypothesis on the bounded character of  $\Omega$  says (after a possible rotation) that for all  $x \in \Omega$  the first component of  $x_1$  is bounded,  $-\infty < a \leq x_1 \leq b < +\infty$ . Let  $v$  be a function of  $C^1(\overline{\Omega})$  which is zero on  $\partial\Omega$ . We can extend it continuously by zero outside of  $\Omega$  ( $v$  is then a continuous function which is piecewise of class  $C^1$  in  $\mathbb{R}^N$ ) and write, for  $x \in \Omega$ ,

$$v(x) = \int_a^{x_1} \frac{\partial v}{\partial x_1}(t, x_2, \dots, x_N) dt,$$

from which we deduce by the Cauchy-Schwarz inequality

$$|v(x)|^2 \leq (x_1 - a) \int_a^{x_1} \left| \frac{\partial v}{\partial x_1}(t, x_2, \dots, x_N) \right|^2 dt \leq (b - a) \int_a^b \left| \frac{\partial v}{\partial x_1}(t, x_2, \dots, x_N) \right|^2 dt.$$

Integrating over  $\Omega$  we obtain

$$\int_{\Omega} |v(x)|^2 dx \leq (b - a) \int_{\Omega} \int_a^b \left| \frac{\partial v}{\partial x_1}(t, x_2, \dots, x_N) \right|^2 dt dx,$$

and permuting the two integrations with respect to  $t$  and  $x$ , we conclude

$$\int_{\Omega} |v(x)|^2 dx \leq (b - a)^2 \int_{\Omega} \left| \frac{\partial v}{\partial x_1}(x) \right|^2 dx \leq (b - a)^2 \int_{\Omega} |\nabla v(x)|^2 dx.$$

□

**Exercise 3.3.1** The aim of this exercise is to show that the space  $V$ , defined by (3.19) and equipped with the scalar product (3.20), is not complete. Let  $\Omega$  be the open unit ball in  $\mathbb{R}^N$ . If  $N = 1$ , we define

$$u_n(x) = \begin{cases} -x - 1 & \text{if } -1 < x < -n^{-1}, \\ (n/2)x^2 - 1 + 1/(2n) & \text{if } -n^{-1} \leq x \leq n^{-1}, \\ x - 1 & \text{if } n^{-1} < x < 1. \end{cases}$$

If  $N = 2$ , for  $0 < \alpha < 1/2$ , we define

$$u_n(x) = |\log(|x|^2/2 + n^{-1})|^\alpha - |\log(1/2 + n^{-1})|^\alpha.$$

If  $N \geq 3$ , for  $0 < \beta < (N - 2)/2$ , we define

$$u_n(x) = \frac{1}{(|x|^2 + n^{-1})^{\beta/2}} - \frac{1}{(1 + n^{-1})^{\beta/2}}.$$

Show that the sequence  $u_n$  is Cauchy in  $V$  but it does not converge in  $V$  as  $n$  tends to infinity.

# 4 Sobolev spaces

---

## 4.1 Introduction and warning

In this chapter we define Sobolev spaces which are **the ‘natural’ spaces of functions in which to solve variational formulations of partial differential equations**. Physically, Sobolev spaces can be interpreted as spaces of **functions with finite energy**. This chapter is the most ‘technical’ of this book and relies in part on a course of ‘pure’ mathematics. Nevertheless, it is necessary to understand the results below well to follow the rest of the course, including its more numerical aspects. In general, it is not necessary to know the proofs of these results (except for the most simple and most useful). However, for the convenience of the reader and to avoid frequent references to other works, we have included most of these proofs. The interested, or curious, reader will find the key ideas and arguments which will allow the understanding of the structure and interest of Sobolev spaces. **Let us emphasize again that it is the spirit of these results more than the details of the proofs which is important here.**

The plan of this chapter is the following. As Sobolev spaces are constructed starting from the idea of a measurable function and from  $L^2$ , the space of square integrable functions, Section 4.2 gives several results on this subject. There we also introduce the idea of **weak differentiation**. Section 4.3 contains all the definitions and the results that we need to know about Sobolev spaces for the remainder of the course. Section 4.4 gives some complementary results for the curious reader. Finally, Section 4.5 allows the reader who knows the theory of distributions (which is not necessary here), to make a link between Sobolev spaces and spaces of distributions. At the end of the chapter **Table 4.1 summarizes all the results that are necessary for what follows.**



## 4.2 Square integrable functions and weak differentiation

### 4.2.1 Some results from integration

All the results of this section are detailed in [4], [35]. Let  $\Omega$  be an open set of  $\mathbb{R}^N$  equipped with the Lebesgue measure. We define by  $L^2(\Omega)$  the space of measurable functions which are square integrable in  $\Omega$ . Under the scalar product

$$\langle f, g \rangle = \int_{\Omega} f(x)g(x) dx,$$

$L^2(\Omega)$  is a Hilbert space (see theorem 3.3.2 of [4]). We denote the corresponding norm by

$$\|f\|_{L^2(\Omega)} = \left( \int_{\Omega} |f(x)|^2 dx \right)^{1/2}.$$

Recall that measurable functions in  $\Omega$  are defined **almost everywhere** in  $\Omega$ : if we change the values of a measurable function  $f$  on a subset of  $\Omega$  of measure zero, we do not change the measurable function  $f$ . In other words, two measurable functions  $f$  and  $g$  are called equal if  $f(x) = g(x)$  almost everywhere in  $\Omega$ , i.e., if there exists  $E \subset \Omega$  such that the Lebesgue measure of  $E$  is zero and  $f(x) = g(x)$  for all  $x \in (\Omega \setminus E)$ .

We denote by  $C_c^\infty(\Omega)$  (or  $\mathcal{D}(\Omega)$ ) the space of functions of class  $C^\infty$  with compact support in  $\Omega$ . We remark that the space  $C_c^\infty(\Omega)$  is not reduced to only the zero function (this is not obvious!, see [4], [38]). Let us note also that the functions of  $C_c^\infty(\Omega)$  are zero, as are all their derivatives, on the boundary of  $\Omega$ . We recall the following density result (see theorem 3.4.3 of [4])

**Theorem 4.2.1** *The space  $C_c^\infty(\Omega)$  is dense in  $L^2(\Omega)$ , that is, for all  $f \in L^2(\Omega)$  there exists a sequence  $f_n \in C_c^\infty(\Omega)$  such that*

$$\lim_{n \rightarrow +\infty} \|f - f_n\|_{L^2(\Omega)} = 0.$$

The following property generalizes lemma 3.2.9.

**Corollary 4.2.2** *Let us take  $f \in L^2(\Omega)$ . If for every function  $\phi \in C_c^\infty(\Omega)$ , we have*

$$\int_{\Omega} f(x)\phi(x) dx = 0,$$

*then  $f(x) = 0$  almost everywhere in  $\Omega$ .*

**Proof.** Let  $f_n \in C_c^\infty(\Omega)$  be the sequence of regular functions which converge to  $f$  in  $L^2(\Omega)$  thanks to theorem 4.2.1. We have

$$0 = \lim_{n \rightarrow +\infty} \int_{\Omega} f(x)f_n(x) dx = \int_{\Omega} |f(x)|^2 dx,$$

from which we deduce that  $f(x) = 0$  almost everywhere in  $\Omega$ . □

More generally, we can define the spaces  $L^p(\Omega)$  with  $1 \leq p \leq +\infty$ . For  $1 \leq p < +\infty$ ,  $L^p(\Omega)$  is the space of measurable functions whose  $p$ th powers are integrable over  $\Omega$ . Equipped with the norm

$$\|f\|_{L^p(\Omega)} = \left( \int_{\Omega} |f(x)|^p dx \right)^{1/p}, \quad (4.1)$$

$L^p(\Omega)$  is a Banach space, that is, a complete normed vector space. For  $p = +\infty$ ,  $L^\infty(\Omega)$  is the space of measurable functions  $f$  which are essentially bounded over  $\Omega$ , that is, there exists a constant  $C > 0$  such that  $|f(x)| \leq C$  almost everywhere in  $\Omega$ . Equipped with the norm

$$\|f\|_{L^\infty(\Omega)} = \inf \{ C \in \mathbb{R}^+ \text{ such that } |f(x)| \leq C \text{ a.e. in } \Omega \}, \quad (4.2)$$

$L^\infty(\Omega)$  is a Banach space. Recall that, if  $\Omega$  is a bounded open set, then  $L^p(\Omega) \subset L^q(\Omega)$  for  $1 \leq q \leq p \leq +\infty$ .

## 4.2.2 Weak differentiation

We first define the concept of the weak derivative in  $L^2(\Omega)$ . This idea generalizes the usual differentiation (sometimes called, by contrast, strong differentiation) and is a particular case of differentiation in the sense of distributions (see Section 4.5 for a brief summary).

**Definition 4.2.3** *Let  $v$  be a function of  $L^2(\Omega)$ . We say that  $v$  is differentiable in the weak sense in  $L^2(\Omega)$  if there exist functions  $w_i \in L^2(\Omega)$ , for  $i \in \{1, \dots, N\}$ , such that, for every function  $\phi \in C_c^\infty(\Omega)$ , we have*

$$\int_{\Omega} v(x) \frac{\partial \phi}{\partial x_i}(x) dx = - \int_{\Omega} w_i(x) \phi(x) dx.$$

*Each  $w_i$  is called the  $i$ th weak partial derivative of  $v$  and is written from now on as  $\frac{\partial v}{\partial x_i}$ .*

Definition 4.2.3 is well defined: in particular, the notation  $w_i = \frac{\partial v}{\partial x_i}$  is unequivocal since, from corollary 4.2.2, the functions  $w_i$  are unique (if they exist). Of course, if  $v$  is differentiable in the usual sense and its partial derivatives belong to  $L^2(\Omega)$ , then the usual and the weak derivatives of  $v$  coincide. Now we give a simple and practical criterion to determine if a function is differentiable in the weak sense.

**Lemma 4.2.4** *Let  $v$  be a function of  $L^2(\Omega)$ . If there exists a constant  $C > 0$  such that, for every function  $\phi \in C_c^\infty(\Omega)$  and for all indices  $i \in \{1, \dots, N\}$ , we have*

$$\left| \int_{\Omega} v(x) \frac{\partial \phi}{\partial x_i}(x) dx \right| \leq C \|\phi\|_{L^2(\Omega)}, \quad (4.3)$$

*then  $v$  is differentiable in the weak sense.*

**Proof.** Let  $L$  be the linear form defined by

$$L(\phi) = \int_{\Omega} v(x) \frac{\partial \phi}{\partial x_i}(x) dx.$$

A priori  $L(\phi)$  is only defined for  $\phi \in C_c^\infty(\Omega)$ , but thanks to the inequality (4.3), we can extend  $L$  by continuity to all functions of  $L^2(\Omega)$  since  $C_c^\infty(\Omega)$  is dense in  $L^2(\Omega)$  from theorem 4.2.1. In fact, the inequality (4.3) proves that the linear form  $L$  is continuous over  $L^2(\Omega)$ . From the Riesz representation theorem 12.1.18, there exists a function  $(-w_i) \in L^2(\Omega)$  such that

$$L(\phi) = - \int_{\Omega} w_i(x) \phi(x) dx,$$

which proves that  $v$  is differentiable in the weak sense in  $L^2(\Omega)$ .  $\square$

**Exercise 4.2.1** Let us take  $\Omega = (0, 1)$ . Show that the function  $x^\alpha$  is differentiable in the weak sense in  $L^2(\Omega)$  if and only if  $\alpha > 1/2$ .

**Exercise 4.2.2** Let  $\Omega$  be an open bounded set. Show that a continuous function over  $\bar{\Omega}$ , which is piecewise  $C^1$ , is differentiable in the weak sense in  $L^2(\Omega)$ .

**Exercise 4.2.3** Let  $\Omega$  be an open bounded set. Show that a piecewise  $C^1$  function which is not continuous is not differentiable in the weak sense in  $L^2(\Omega)$ .

We recover a well-known result for the usual derivative.

**Proposition 4.2.5** *Let  $v$  be a function of  $L^2(\Omega)$  which is differentiable in the weak sense and such that all its weak partial derivatives  $\frac{\partial v}{\partial x_i}$ , for  $1 \leq i \leq N$ , are zero. Then, for every connected component  $\Omega$ , there exists a constant  $C$  such that  $v(x) = C$  almost everywhere in this connected component.*

**Proof.** For all  $\psi \in C_c^\infty(\Omega)$ , we have

$$\int_{\Omega} v(x) \frac{\partial \psi}{\partial x_i}(x) dx = 0. \quad (4.4)$$

Let  $Q = ]-\ell, +\ell[^N$  be an open cube contained in  $\Omega$  (with  $\ell > 0$ ), and let  $\theta(t) \in C_c^\infty(-\ell, +\ell)$  be such that

$$\int_{-\ell}^{+\ell} \theta(t) dt = 1.$$

For every function  $\phi \in C_c^\infty(Q)$  we define

$$\psi(x', x_i) = \int_{-\ell}^{x_i} \left( \theta(t) \int_{-\ell}^{+\ell} \phi(x', s) ds - \phi(x', t) \right) dt,$$

with the notation  $x = (x', x_i)$  for  $x' \in \mathbb{R}^{N-1}$  and  $x_i \in \mathbb{R}$ . We easily verify that  $\psi$  also belongs to  $C_c^\infty(Q)$  and that

$$\frac{\partial \psi}{\partial x_i}(x', x_i) = \theta(x_i) \int_{-\ell}^{+\ell} \phi(x', s) ds - \phi(x', x_i).$$

With such a function  $\psi$  the equation (4.4) becomes

$$\begin{aligned} \int_Q v(x) \phi(x) dx &= \int_Q v(x) \theta(x_i) \left( \int_{-\ell}^{+\ell} \phi(x', s) ds \right) dx' dx_i \\ &= \int_Q \phi(x', s) \left( \int_{-\ell}^{+\ell} v(x', x_i) \theta(x_i) dx_i \right) dx' ds \end{aligned}$$

thanks to the Fubini theorem. As  $\phi$  is arbitrary, by application of corollary 4.2.2 we deduce

$$v(x) = \int_{-\ell}^{+\ell} v(x', s) \theta(s) ds,$$

that is,  $v$  does not depend on  $x_i$  in  $Q$ . By repeating this argument for all components  $x_i$ , we find that  $v(x)$  is constant in  $Q$ . Since any pair of points in the same connected component of  $\Omega$  can be linked by a chain of such cubes (of variable size) we can conclude that  $v(x)$  is constant in every connected component of  $\Omega$ .  $\square$

We can easily generalize the definition of the weak derivative to differential operators which only involve some (but not all) combinations of partial derivatives. This is the case, for example, for the divergence of a vector valued function which we shall use later.

**Definition 4.2.6** Let  $\sigma$  be a function from  $\Omega$  into  $\mathbb{R}^N$  with all components belonging to  $L^2(\Omega)$  (we say  $\sigma \in L^2(\Omega)^N$ ). We say that  $\sigma$  has divergence in the weak sense in  $L^2(\Omega)$  if there exists a function  $w \in L^2(\Omega)$  such that, for every function  $\phi \in C_c^\infty(\Omega)$ , we have

$$\int_{\Omega} \sigma(x) \cdot \nabla \phi(x) dx = - \int_{\Omega} w(x) \phi(x) dx.$$

The function  $w$  is called the weak divergence of  $\sigma$  and from now on will be denoted as  $\text{div} \sigma$ .

The justification of definition 4.2.6 is that, if  $\sigma$  is a regular function, then a simple integration by parts (see corollary 3.2.3) shows that we have  $w = \text{div} \sigma$ . An easy generalization of this criterion of weak differentiation in lemma 4.2.4 is given by the following result (whose proof we leave to the reader as an exercise).

**Lemma 4.2.7** Let  $\sigma$  be a function of  $L^2(\Omega)^N$ . If there exists a constant  $C > 0$  such that, for every function  $\phi \in C_c^\infty(\Omega)$ , we have

$$\left| \int_{\Omega} \sigma(x) \cdot \nabla \phi(x) dx \right| \leq C \|\phi\|_{L^2(\Omega)},$$

then  $\sigma$  has a divergence in the weak sense.

**Exercise 4.2.4** Let  $\Omega$  be an open bounded set composed of two open sets  $\Omega_1$  and  $\Omega_2$  separated by a surface  $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ . Show that a vector function of class  $C^1$  over each part  $\Omega_1$  and  $\Omega_2$  has a weak divergence in  $L^2(\Omega)$  if and only if its normal component is continuous across the surface  $\Gamma$ .

**Remark 4.2.8** The idea of weak differentiation and all the results of this subsection extend to spaces  $L^p(\Omega)$  for  $1 \leq p \leq +\infty$ . Since for  $p \neq 2$ , the space  $L^p(\Omega)$  is not a Hilbert space, the criterion of weak differentiation (4.3) must be replaced by

$$\left| \int_{\Omega} v(x) \frac{\partial \phi}{\partial x_i}(x) dx \right| \leq C \|\phi\|_{L^{p'}(\Omega)} \quad \text{with } \frac{1}{p} + \frac{1}{p'} = 1 \text{ and } 1 < p \leq +\infty,$$

and the Riesz representation theorem 12.1.18 is replaced by using the dual  $L^p(\Omega)$  of  $L^{p'}(\Omega)$  (see [6]). •

## 4.3 Definition and principal properties

### 4.3.1 The space $H^1(\Omega)$

**Definition 4.3.1** Let  $\Omega$  be an open set of  $\mathbb{R}^N$ . The Sobolev space  $H^1(\Omega)$  is defined by

$$H^1(\Omega) = \left\{ v \in L^2(\Omega) \text{ such that } \forall i \in \{1, \dots, N\} \frac{\partial v}{\partial x_i} \in L^2(\Omega) \right\}, \quad (4.5)$$

where  $\frac{\partial v}{\partial x_i}$  is the weak partial derivative of  $v$  in the sense of definition 4.2.3.

In physics and mechanics the Sobolev space is often called the **energy space** in the sense that it is composed of functions with finite energy (that is, the norm  $\|u\|_{H^1(\Omega)}$  is finite). Functions of finite energy can possibly be ‘singular’ which has a possible physical sense (a concentration or a localized explosion). We shall look with interest at the explicit examples of exercise 4.3.2 and of lemma 5.2.33.

**Proposition 4.3.2** Equipped with the scalar product

$$\langle u, v \rangle = \int_{\Omega} \left( u(x)v(x) + \nabla u(x) \cdot \nabla v(x) \right) dx \quad (4.6)$$

and with the norm

$$\|u\|_{H^1(\Omega)} = \left( \int_{\Omega} (|u(x)|^2 + |\nabla u(x)|^2) dx \right)^{1/2}$$

the Sobolev space  $H^1(\Omega)$  is a Hilbert space.

**Proof.** It is obvious that (4.6) is a scalar product in  $H^1(\Omega)$ . It, therefore, remains to show that  $H^1(\Omega)$  is complete for the associated norm. Let  $(u_n)_{n \geq 1}$  be a Cauchy sequence in  $H^1(\Omega)$ . By definition of the norm  $H^1(\Omega)$ ,  $(u_n)_{n \geq 1}$  as well as  $(\frac{\partial u_n}{\partial x_i})_{n \geq 1}$  for  $i \in \{1, \dots, N\}$  are Cauchy sequences in  $L^2(\Omega)$ . As  $L^2(\Omega)$  is complete, there exist limits  $u$  and  $w_i$  such that  $u_n$  converges to  $u$  and  $\frac{\partial u_n}{\partial x_i}$  converges to  $w_i$  in  $L^2(\Omega)$ . Now, by definition of the weak derivative of  $u_n$ , for every function  $\phi \in C_c^\infty(\Omega)$ , we have

$$\int_{\Omega} u_n(x) \frac{\partial \phi}{\partial x_i}(x) dx = - \int_{\Omega} \frac{\partial u_n}{\partial x_i}(x) \phi(x) dx. \quad (4.7)$$

Passing to the limit  $n \rightarrow +\infty$  in (4.7), we obtain

$$\int_{\Omega} u(x) \frac{\partial \phi}{\partial x_i}(x) dx = - \int_{\Omega} w_i(x) \phi(x) dx,$$

which proves that  $u$  is differentiable in the weak sense and that  $w_i$  is the  $i$ th weak partial derivative of  $u$ ,  $\frac{\partial u}{\partial x_i}$ . Therefore,  $u$  belongs to  $H^1(\Omega)$  and  $(u_n)_{n \geq 1}$  converges to  $u$  in  $H^1(\Omega)$ .  $\square$

**Exercise 4.3.1** Show that piecewise  $C^1$ , continuous functions, with bounded support in  $\overline{\Omega}$ , belong to  $H^1(\Omega)$ .

In  $N \geq 2$  dimensions, the functions of  $H^1(\Omega)$  are in general **neither continuous nor bounded**, as the following counterexample shows.

**Exercise 4.3.2** Let  $B$  be the open unit ball of  $\mathbb{R}^N$ . If  $N = 2$ , show that the function  $u(x) = |\log(|x|/2)|^\alpha$  belongs to  $H^1(B)$  for  $0 < \alpha < 1/2$ , but is not bounded in the neighbourhood of the origin. If  $N \geq 3$ , show that the function  $u(x) = |x|^{-\beta}$  belongs to  $H^1(B)$  for  $0 < \beta < (N-2)/2$ , but is not bounded in the neighbourhood of the origin.

The space of dimension  $N = 1$  is an ‘exception’ to the noncontinuity of functions of  $H^1(\Omega)$  as we show in the following lemma where, without loss of generality, we take  $\Omega = (0, 1)$ .

**Lemma 4.3.3** For every function  $v \in H^1(0, 1)$  and for all  $x, y \in [0, 1]$ , we have

$$v(y) = v(x) + \int_x^y v'(s) ds. \quad (4.8)$$

More generally, for all  $x \in [0, 1]$ , the mapping  $v \rightarrow v(x)$ , defined from  $H^1(0, 1)$  into  $\mathbb{R}$ , is a continuous linear form over  $H^1(0, 1)$ . In particular, every function  $v \in H^1(0, 1)$  is continuous over  $[0, 1]$ .

**Proof.** Let us take  $v \in H^1(0, 1)$ . We define a function  $w(x)$  over  $[0, 1]$  by

$$w(x) = \int_0^x v'(s) ds.$$

This definition has a sense as, by the Cauchy–Schwarz inequality,

$$\left| \int_0^x v'(s) ds \right| \leq \sqrt{x} \sqrt{\int_0^x |v'(s)|^2 ds} \leq \sqrt{\int_0^1 |v'(s)|^2 ds} < +\infty.$$

In fact, the same argument shows that the function  $w$  is continuous over  $[0, 1]$

$$|w(x) - w(y)| = \left| \int_y^x v'(s) ds \right| \leq \sqrt{|x - y|} \sqrt{\int_y^x |v'(s)|^2 ds} \leq \sqrt{|x - y|} \sqrt{\int_0^1 |v'(s)|^2 ds}.$$

We show that  $w$  is differentiable in the weak sense and that  $w' = v'$ . Let  $\phi \in C_c^\infty(0, 1)$ . Denoting by  $T$  the triangle  $T = \{(x, s) \in \mathbb{R}^2, 0 \leq s \leq x \leq 1\}$ , we have

$$\int_0^1 w(x) \phi'(x) dx = \int_0^1 \left( \int_0^x v'(s) ds \right) \phi'(x) dx = \int_T v'(s) \phi'(x) ds dx.$$

By application of the Fubini theorem, we have

$$\int_T v'(s) \phi'(x) ds dx = \int_0^1 \left( \int_s^1 \phi'(x) dx \right) v'(s) ds = - \int_0^1 \phi(s) v'(s) ds,$$

and by Cauchy–Schwarz we deduce

$$\left| \int_0^1 w(x) \phi'(x) dx \right| \leq \|v'\|_{L^2(0,1)} \|\phi\|_{L^2(0,1)}.$$

Therefore,  $w$  is differentiable in the weak sense and by the definition of  $w'$  we have

$$- \int_0^1 w'(x) \phi(x) dx = \int_0^1 w(x) \phi'(x) dx = - \int_0^1 \phi(s) v'(s) ds,$$

for all  $\phi \in C_c^\infty(0, 1)$ , which implies that  $w' = v'$ . Lemma 4.2.5 tells us that  $w - v$  is equal to a constant almost everywhere in  $(0, 1)$ , which establishes (4.8). Starting from (4.8) and by using Cauchy–Schwarz we obtain

$$|v(x)| \leq |v(y)| + \sqrt{|y - x|} \sqrt{\int_x^y |v'(s)|^2 ds} \leq |v(y)| + \sqrt{\int_0^1 |v'(s)|^2 ds},$$

and integrating with respect to  $y$

$$\begin{aligned} |v(x)| &\leq \int_0^1 |v(y)| dy + \sqrt{\int_0^1 |v'(s)|^2 ds} \\ &\leq \sqrt{\int_0^1 |v(y)|^2 dy} + \sqrt{\int_0^1 |v'(s)|^2 ds} \leq \sqrt{2} \|v\|_{H^1(0,1)}, \end{aligned}$$

which proves that  $v \rightarrow v(x)$  is a continuous linear form over  $H^1(0, 1)$ .  $\square$

**Remark 4.3.4** The assertion that every function of  $H^1(0,1)$  is continuous may seem, at first glance, contradictory to the fact that the functions of  $H^1(0,1)$ , are all measurable functions, only defined almost everywhere (in other words, we can change some point values of  $v$  without changing the function in  $H^1(0,1)$  and not destroying the continuity of  $v$ ). To solve this apparent paradox, we must recall that a function of  $H^1(0,1)$  is in fact a class of functions since we identify two functions that are equal almost everywhere. Under these conditions, the result of lemma 4.3.3 must be understood in the sense that there exists a representative of the class of functions  $v \in H^1(0,1)$  which is continuous. •

It is very important in practice to know if **regular functions are dense in the Sobolev space  $H^1(\Omega)$** . This partly justifies the idea of a Sobolev space which occurs very simply as the set of regular functions completed by the limits of sequences of regular functions in the energy norm  $\|u\|_{H^1(\Omega)}$ . This allows us to prove several properties easily by establishing them first for regular functions then by using a ‘density’ argument (see, for example, the proofs of theorems 4.3.13 and 4.3.15).

**Theorem 4.3.5 (density)** *If  $\Omega$  is a regular open bounded set of class  $C^1$ , or if  $\Omega = \mathbb{R}_+^N$ , or even if  $\Omega = \mathbb{R}^N$ , then  $C_c^\infty(\bar{\Omega})$  is dense in  $H^1(\Omega)$ .*

The proof of theorem 4.3.5 is found in Section 4.4. We recall that the notation  $\mathbb{R}_+^N$  denotes the half-space  $\{x \in \mathbb{R}^N \text{ such that } x_N > 0\}$ .

**Remark 4.3.6** The space  $C_c^\infty(\bar{\Omega})$  which is dense in  $H^1(\Omega)$  is composed of regular functions of class  $C^\infty$  with bounded (or compact) support in the closed set  $\bar{\Omega}$ . In particular, if  $\Omega$  is bounded, all the functions of  $C_c^\infty(\bar{\Omega})$  necessarily have bounded support, and therefore  $C_c^\infty(\bar{\Omega}) = C^\infty(\bar{\Omega})$ . We point out that functions of  $C_c^\infty(\bar{\Omega})$  are not necessarily zero on the boundary of the open set  $\Omega$ , which is what differentiates this space from  $C_c^\infty(\Omega)$  (see remark 3.2.2). Conversely, if  $\Omega$  is not bounded, the  $C_c^\infty(\bar{\Omega})$  functions are zero ‘at infinity’. •

**Remark 4.3.7** The concept of regularity of an open set has been introduced in definition 3.2.5. It is not necessary to know the precise details of this definition of the regularity of an open set. It is sufficient to know *roughly speaking* that we need the boundary of the open set to be a regular surface and that we exclude certain ‘pathologies’ (see remark 3.2.6). When we state a result under a regularity hypothesis on the open set, this regularity is always necessary (the result fails for certain nonregular open sets, see the counterexample of exercise 4.3.3). Nevertheless, the regularity hypothesis in definition 3.2.5 can often be weakened: the membership of a class of functions  $C^1$  can be replaced by the membership of the class of Lipschitz functions (see [14]). Although these details are largely beyond the scope of this course, we make this remark so that the careful reader does not object when we use such results (where the regularity hypothesis is necessary) in the case of open sets ‘with corners’ which appear naturally in all numerical calculations (see the different grids which illustrate this course). •



### 4.3.2 The space $H_0^1(\Omega)$

Let us now define another Sobolev space which is a subspace of  $H^1(\Omega)$  and which will be very useful for problems with Dirichlet boundary conditions.

**Definition 4.3.8** Let  $C_c^\infty(\Omega)$  be the space of functions of class  $C^\infty$  with compact support in  $\Omega$ . The Sobolev space  $H_0^1(\Omega)$  is defined as the closure of  $C_c^\infty(\Omega)$  in  $H^1(\Omega)$ .

We shall see a little later (see corollary 4.3.16) that  $H_0^1(\Omega)$  is in fact the subspace of  $H^1(\Omega)$  composed of **functions which are zero on the boundary**  $\partial\Omega$  since this is the case for functions of  $C_c^\infty(\Omega)$ . In general,  $H_0^1(\Omega)$  is **strictly smaller** than  $H^1(\Omega)$  since  $C_c^\infty(\Omega)$  is a **strict** subspace of  $C_c^\infty(\bar{\Omega})$  (see theorem 4.3.5 and remark 4.3.6). An important exception is the case where  $\Omega = \mathbb{R}^N$ : in effect, in this case  $\bar{\Omega} = \mathbb{R}^N = \Omega$  and theorem 4.3.5 shows that  $C_c^\infty(\mathbb{R}^N)$  is dense in  $H^1(\mathbb{R}^N)$ , therefore we have  $H_0^1(\mathbb{R}^N) = H^1(\mathbb{R}^N)$ . This exception is easily understood as the whole space  $\mathbb{R}^N$  does not have a boundary.

**Proposition 4.3.9** Equipped with the scalar product (4.6) of  $H^1(\Omega)$ , the Sobolev space  $H_0^1(\Omega)$  is a Hilbert space.

**Proof.** By definition  $H_0^1(\Omega)$  is a closed subspace of  $H^1(\Omega)$  (which is a Hilbert space), therefore it is also a Hilbert space.  $\square$

An essential result for the applications of the next chapter is the following inequality.

**Proposition 4.3.10 (Poincaré inequality)** Let  $\Omega$  be an open set of  $\mathbb{R}^N$  which is bounded in at least one space direction. There exists a constant  $C > 0$  such that, for every function  $v \in H_0^1(\Omega)$ ,

$$\int_{\Omega} |v(x)|^2 dx \leq C \int_{\Omega} |\nabla v(x)|^2 dx. \quad (4.9)$$

**Proof.** For functions  $v \in C_c^\infty(\Omega)$  we have already proved the Poincaré inequality (4.9) in lemma 3.3.6. By a density argument the result remains true for every function  $v \in H_0^1(\Omega)$ . In effect, as  $C_c^\infty(\Omega)$  is dense in  $H_0^1(\Omega)$  (by definition 4.3.8), there exists a sequence  $v_n \in C_c^\infty(\Omega)$  such that

$$\lim_{n \rightarrow +\infty} \|v - v_n\|_{H^1(\Omega)}^2 = \lim_{n \rightarrow +\infty} \int_{\Omega} (|v - v_n|^2 + |\nabla(v - v_n)|^2) dx = 0.$$

In particular, we deduce that

$$\lim_{n \rightarrow +\infty} \int_{\Omega} |v_n|^2 dx = \int_{\Omega} |v|^2 dx \quad \text{and} \quad \lim_{n \rightarrow +\infty} \int_{\Omega} |\nabla v_n|^2 dx = \int_{\Omega} |\nabla v|^2 dx.$$

By applying the lemma 3.3.6, we have

$$\int_{\Omega} |v_n(x)|^2 dx \leq C \int_{\Omega} |\nabla v_n(x)|^2 dx. \quad (4.10)$$

We then pass to the limit  $n \rightarrow +\infty$  in each of the two terms of the inequality (4.10) to obtain the result. This type of ‘density’ argument will be used frequently from now on.  $\square$

**Remark 4.3.11** The Poincaré inequality (4.9) is not true for functions of  $H^1(\Omega)$ . In effect, the constant (nonzero) functions make the term on the right of (4.9) zero but not the term on the left. The essential hypothesis in the Poincaré inequality is that the functions of  $H_0^1(\Omega)$  are zero on the boundary  $\partial\Omega$  of the open set  $\Omega$  (see remark 4.3.18 for some variants of this hypothesis).  $\bullet$

An important corollary of the Poincaré inequality is the following result which gives a simpler equivalent norm in  $H_0^1(\Omega)$ .

**Corollary 4.3.12** *Let  $\Omega$  be an open set of  $\mathbb{R}^N$  bounded in at least one space direction. Then the seminorm*

$$|v|_{H_0^1(\Omega)} = \left( \int_{\Omega} |\nabla v(x)|^2 dx \right)^{1/2}$$

*is a norm over  $H_0^1(\Omega)$  which is equivalent to the usual norm induced by that of  $H^1(\Omega)$ .*

**Proof.** Take  $v \in H_0^1(\Omega)$ . The first inequality

$$|v|_{H_0^1(\Omega)} \leq \|v\|_{H^1(\Omega)} = \left( \int_{\Omega} (|v|^2 + |\nabla v|^2) dx \right)^{1/2}$$

is obvious. On the other hand, the Poincaré inequality of lemma 3.3.6 leads to

$$\|v\|_{H^1(\Omega)}^2 \leq (C+1) \int_{\Omega} |\nabla v|^2 dx = (C+1) |v|_{H_0^1(\Omega)}^2,$$

which proves that  $|v|_{H_0^1(\Omega)}$  is a norm equivalent to  $\|v\|_{H^1(\Omega)}$ .  $\square$

### 4.3.3 Traces and Green’s formulas

We have seen that, in  $N \geq 2$  dimensions, the functions of  $H^1(\Omega)$  are usually not continuous (see the counterexample of exercise 4.3.2). Therefore, as for every measurable function, we cannot speak of the pointwise value of a function  $v \in H^1(\Omega)$  but only ‘almost everywhere’ in  $\Omega$ . In particular, it is not obvious if we can define the ‘boundary value’, or ‘trace’ of  $v$  on the boundary  $\partial\Omega$  since  $\partial\Omega$  is a set of measure zero. Very fortunately for boundary value problems which we study, it is possible to define the trace  $v|_{\partial\Omega}$  of a function of  $H^1(\Omega)$ . This essential result, called the trace theorem, is the following.

**Theorem 4.3.13 (trace)** *Let  $\Omega$  be an open bounded regular set of class  $\mathcal{C}^1$ , or  $\Omega = \mathbb{R}_+^N$ . We define the trace mapping  $\gamma_0$*

$$\begin{aligned} H^1(\Omega) \cap C(\overline{\Omega}) &\rightarrow L^2(\partial\Omega) \cap C(\overline{\partial\Omega}) \\ v &\rightarrow \gamma_0(v) = v|_{\partial\Omega}. \end{aligned} \quad (4.11)$$

*This mapping  $\gamma_0$  is extended by continuity to a continuous linear mapping of  $H^1(\Omega)$  into  $L^2(\partial\Omega)$ , again called  $\gamma_0$ . In particular, there exists a constant  $C > 0$  such that, for every function  $v \in H^1(\Omega)$ , we have*

$$\|v\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)}. \quad (4.12)$$

**Remark 4.3.14** Thanks to the trace theorem (4.3.13), we can talk of the value of a function of  $H^1(\Omega)$  on the boundary  $\partial\Omega$ . This result is remarkable as it is not true for a function of  $L^2(\Omega)$  (see in particular exercise 4.3.4). •

**Proof.** Let us prove the result for the half-space  $\Omega = \mathbb{R}_+^N = \{x \in \mathbb{R}^N, x_N > 0\}$ . Let  $v \in C_c^\infty(\overline{\mathbb{R}_+^N})$ . With the notation  $x = (x', x_N)$ , we have

$$|v(x', 0)|^2 = -2 \int_0^{+\infty} v(x', x_N) \frac{\partial v}{\partial x_N}(x', x_N) dx_N,$$

and, by using the inequality  $2ab \leq a^2 + b^2$ ,

$$|v(x', 0)|^2 \leq \int_0^{+\infty} \left( |v(x', x_N)|^2 + \left| \frac{\partial v}{\partial x_N}(x', x_N) \right|^2 \right) dx_N.$$

By integration in  $x'$ , we deduce

$$\int_{\mathbb{R}^{N-1}} |v(x', 0)|^2 dx' \leq \int_{\mathbb{R}_+^N} \left( |v(x)|^2 + \left| \frac{\partial v}{\partial x_N}(x) \right|^2 \right) dx,$$

that is,  $\|v\|_{L^2(\partial\mathbb{R}_+^N)} \leq \|v\|_{H^1(\mathbb{R}_+^N)}$ . By the density of  $C_c^\infty(\overline{\mathbb{R}_+^N})$  in  $H^1(\mathbb{R}_+^N)$ , we therefore obtain the result.

For an open bounded regular set of class  $\mathcal{C}^1$ , we use an argument involving local coordinates on the boundary which allows us to reduce it to the case of  $\Omega = \mathbb{R}_+^N$ . We do not detail this argument (which is too technical) which is the same as that used in the proof of proposition 4.4.2. □

The trace theorem 4.3.13 allows us to generalize, to functions of  $H^1(\Omega)$ , the Green's formula which has been established for functions of class  $\mathcal{C}^1$  in corollary 3.2.3.

**Theorem 4.3.15 (Green's formula)** *Let  $\Omega$  be an open bounded regular set of class  $\mathcal{C}^1$ . If  $u$  and  $v$  are functions of  $H^1(\Omega)$ , they satisfy*

$$\int_{\Omega} u(x) \frac{\partial v}{\partial x_i}(x) dx = - \int_{\Omega} v(x) \frac{\partial u}{\partial x_i}(x) dx + \int_{\partial\Omega} u(x)v(x)n_i(x) ds, \quad (4.13)$$

where  $n = (n_i)_{1 \leq i \leq N}$  is the outward unit normal to  $\partial\Omega$ .

**Proof.** Recall that the formula (4.13) has been established for functions of class  $C^1$  in the corollary 3.2.3. We again use a density argument. By the density of  $C_c^\infty(\bar{\Omega})$  in  $H^1(\Omega)$  (see theorem 4.3.5), there exist sequences  $(u_n)_{n \geq 1}$  and  $(v_n)_{n \geq 1}$  in  $C_c^\infty(\bar{\Omega})$  which converge in  $H^1(\Omega)$  to  $u$  and  $v$ , respectively. From Corollary 3.2.3 we have

$$\int_{\Omega} u_n \frac{\partial v_n}{\partial x_i} dx = - \int_{\Omega} v_n \frac{\partial u_n}{\partial x_i} dx + \int_{\partial\Omega} u_n v_n n_i ds. \quad (4.14)$$

We can pass to the limit  $n \rightarrow +\infty$  in the first two terms of (4.14) as  $u_n$  and  $\frac{\partial u_n}{\partial x_i}$  (respectively,  $v_n$  and  $\frac{\partial v_n}{\partial x_i}$ ) converge to  $u$  and  $\frac{\partial u}{\partial x_i}$  (respectively,  $v$  and  $\frac{\partial v}{\partial x_i}$ ) in  $L^2(\Omega)$ . To pass to the limit in the last integral of (4.14), we use the continuity of the trace mapping  $\gamma_0$ , that is, the inequality (4.12), which allows us to check that  $\gamma_0(u_n)$  (respectively,  $\gamma_0(v_n)$ ) converges to  $\gamma_0(u)$  (respectively,  $\gamma_0(v)$ ) in  $L^2(\partial\Omega)$ . We therefore obtain the formula (4.13) for functions  $u$  and  $v$  of  $H^1(\Omega)$ .  $\square$

As a consequence of the trace theorem 4.3.13 we obtain a very simple characterization of the space  $H_0^1(\Omega)$ .

**Corollary 4.3.16** *Let  $\Omega$  be an open bounded regular set of class  $C^1$ . The space  $H_0^1(\Omega)$  coincides with the subspace of  $H^1(\Omega)$  composed of functions which are zero on the boundary  $\partial\Omega$ .*

**Proof.** As every function of  $H_0^1(\Omega)$  is the limit of a sequence of functions belonging to  $C_c^\infty(\Omega)$  which have zero trace, the continuity of the trace mapping  $\gamma_0$  implies that the trace of the limit is also zero. We deduce that  $H_0^1(\Omega)$  is contained in the subspace of  $H^1(\Omega)$  of functions which are zero on the boundary  $\partial\Omega$ . The reciprocal is more technical and follows from a double procedure of local coordinates (see the proof of proposition 4.4.2) then of regularization and translation (similar to the proof of theorem 4.4.1). We refer to [6], [34] for more details.  $\square$

**Remark 4.3.17** Corollary 4.3.16 confirms that the kernel of the trace mapping  $\gamma_0$  is exactly  $H_0^1(\Omega)$ . A natural, but delicate, question is to characterize the image of  $\gamma_0$ . We shall be content with saying that this image  $\text{Im}(\gamma_0)$  is not equal to  $L^2(\partial\Omega)$ , but a strict subspace, which is dense in  $L^2(\partial\Omega)$ , composed of ‘more regular’ functions, and denoted  $H^{1/2}(\partial\Omega)$ . For more details, we refer to [28].  $\bullet$

Thanks to corollary 4.3.16 we can give another proof of proposition 4.3.10 using the Poincaré inequality. This new proof is no longer ‘constructive’ but is based on an argument by contradiction which has the merit that it can be generalized very easily. In effect, there exist numerous variants of the Poincaré inequality, adapted to different models of partial differential equations. In view of the importance of this inequality for what follows, we shall give a proof which is easily adaptable to all the cases that occur.

**Another proof of proposition 4.3.10.** We proceed by contradiction. If there does not exist a  $C > 0$  such that, for every function  $v \in H_0^1(\Omega)$ ,

$$\int_{\Omega} |v(x)|^2 dx \leq C \int_{\Omega} |\nabla v(x)|^2 dx,$$

this means that there exists a sequence  $v_n \in H_0^1(\Omega)$  such that

$$1 = \int_{\Omega} |v_n(x)|^2 dx > n \int_{\Omega} |\nabla v_n(x)|^2 dx. \quad (4.15)$$

In particular, (4.15) implies that the sequence  $v_n$  is bounded in  $H_0^1(\Omega)$ . By an application of the Rellich theorem 4.3.21, there exists a subsequence  $v_{n'}$  which converges in  $L^2(\Omega)$ . Further, (4.15) shows that the sequence  $\nabla v_{n'}$  converges to zero in  $L^2(\Omega)$  (component by component). Consequently,  $v_{n'}$  is a Cauchy sequence in  $H_0^1(\Omega)$ , which is a Hilbert space, therefore, it converges in  $H_0^1(\Omega)$  to a limit  $v$ . As we have

$$\int_{\Omega} |\nabla v(x)|^2 dx = \lim_{n \rightarrow +\infty} \int_{\Omega} |\nabla v_n(x)|^2 dx \leq \lim_{n \rightarrow +\infty} \frac{1}{n} = 0,$$

we deduce, from lemma 4.2.5, that  $v$  is a constant in each connected component of  $\Omega$ . But as  $v$  is zero on the boundary  $\partial\Omega$  (from corollary 4.3.16),  $v$  is identically zero in all  $\Omega$ . In addition,

$$\int_{\Omega} |v(x)|^2 dx = \lim_{n \rightarrow +\infty} \int_{\Omega} |v_n(x)|^2 dx = 1,$$

which is a contradiction with the fact that  $v = 0$ . □

**Remark 4.3.18** The proof by contradiction of proposition 4.3.10 easily generalizes. We take, for example, the case of an open, bounded connected set  $\Omega$ , which is regular of class  $\mathcal{C}^1$ , and whose boundary  $\partial\Omega$  decomposes into two disjoint regular parts  $\partial\Omega_N$  and  $\partial\Omega_D$  whose surface measures are nonzero (see Figure 4.1). We define a space  $V$  by

$$V = \{v \in H^1(\Omega) \text{ such that } v = 0 \text{ on } \partial\Omega_D\}.$$

By application of the trace theorem 4.3.13, it is easy to see that  $V$  is a closed subspace of  $H^1(\Omega)$ , and therefore is a Hilbert space for the scalar product of  $H^1(\Omega)$ . As for  $H_0^1(\Omega)$ , the argument by contradiction allows us to prove the existence of a constant  $C > 0$  such that every function  $v \in V$  satisfies the Poincaré inequality (4.9). •

By application of Green's formula of theorem 4.3.15, we can construct a family of examples of functions belonging to  $H^1(\Omega)$ . This family of examples will be very useful to us in what follows to construct finite dimensional subspaces of  $H^1(\Omega)$ .

**Lemma 4.3.19** *Let  $\Omega$  be an open bounded regular set of class  $\mathcal{C}^1$ . Let  $(\omega_i)_{1 \leq i \leq I}$  be a regular partition of  $\Omega$ , that is, each  $\omega_i$  is a regular open set of class  $\mathcal{C}^1$ ,  $\omega_i \cap \omega_j = \emptyset$  if  $i \neq j$ , and  $\overline{\Omega} = \cup_{i=1}^I \overline{\omega_i}$ . Let  $v$  be a function whose restriction to each  $\omega_i$ ,  $v_i = v|_{\omega_i}$ , belongs to  $H^1(\omega_i)$ . If  $v$  is continuous over  $\overline{\Omega}$ , then  $v$  belongs to  $H^1(\Omega)$ .*

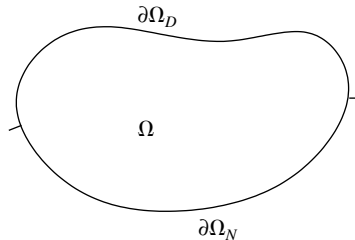


Figure 4.1. Partition in two disjoint parts of the boundary of an open set.

**Proof.** Let us calculate the weak derivative of  $v$ : for  $\phi \in C_c^\infty(\Omega)$ , by application of Green's formula in each  $\omega_i$  we have

$$\begin{aligned} \int_{\Omega} v(x) \frac{\partial \phi}{\partial x_j}(x) dx &= \sum_{i=1}^I \int_{\omega_i} v_i(x) \frac{\partial \phi}{\partial x_j}(x) dx \\ &= - \sum_{i=1}^I \int_{\omega_i} \frac{\partial v_i}{\partial x_j}(x) \phi(x) dx + \sum_{i=1}^I \int_{\partial \omega_i} v_i(x) \phi(x) n_j^i(x) ds \\ &= - \sum_{i=1}^I \int_{\omega_i} \frac{\partial v_i}{\partial x_j}(x) \phi(x) dx, \end{aligned}$$

since the integrals on the boundary cancel. In effect, on the part  $\Gamma = \partial \omega_i \cap \partial \omega_k$  of the boundary shared by the two open sets  $\omega_i$  and  $\omega_k$ , we have  $n_j^i(x) = -n_j^k(x)$  and therefore, by continuity of  $v$  and  $\phi$ ,

$$\int_{\Gamma} v_i(x) \phi(x) n_j^i(x) ds + \int_{\Gamma} v_k(x) \phi(x) n_j^k(x) ds = 0.$$

We deduce therefore that  $v$  is differentiable in the weak sense and that

$$\left. \frac{\partial v}{\partial x_j} \right|_{\omega_i} = \frac{\partial v_i}{\partial x_j}.$$

In particular, this implies that  $v$  belongs to  $H^1(\Omega)$ . □

**Remark 4.3.20** It is not necessarily easy to decompose a regular open set  $\Omega$  into a partition of regular open sets  $(\omega_i)_{1 \leq i \leq I}$ . Fortunately, lemma 4.3.19 remains true if the open sets  $\omega_i$  are only 'piecewise' regular. We shall sometimes use this very convenient generalization of lemma 4.3.19. •

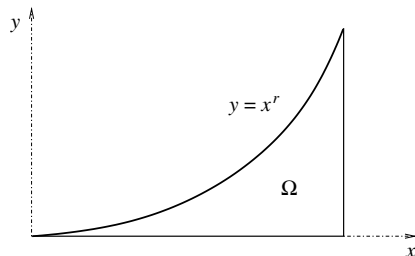


Figure 4.2. Example of a nonregular open set.

**Exercise 4.3.3** The aim of this exercise is to show that the trace theorem 4.3.13 is not true if the open set  $\Omega$  is not regular. Take the open set  $\Omega \subset \mathbb{R}^2$  defined by  $0 < x < 1$  and  $0 < y < x^r$  with  $r > 2$  (see Figure 4.2). Take the function  $v(x) = x^\alpha$ . Show that  $v \in H^1(\Omega)$  if and only if  $2\alpha + r > 1$ , while  $v \in L^2(\partial\Omega)$  if and only if  $2\alpha > -1$ . (We can also show with this same example that the density theorem 4.3.5 and prolongation proposition 4.4.2 are not true for such an open set.) Conclude the result.

**Exercise 4.3.4** The aim of this exercise is to show that we cannot have the idea of trace for functions of  $L^2(\Omega)$ , that is, there does not exist a constant  $C > 0$  such that, for every function  $v \in L^2(\Omega)$ , we have

$$\|v|_{\partial\Omega}\|_{L^2(\partial\Omega)} \leq C\|v\|_{L^2(\Omega)}.$$

For simplicity, we choose as the set  $\Omega$  the unit ball. Construct a sequence of regular functions in  $\bar{\Omega}$  equal to 1 on  $\partial\Omega$  and whose norm in  $L^2(\Omega)$  tends to zero. Conclude the result.

#### 4.3.4 A compactness result

We devote this section to the study of a compactness property known as the Rellich theorem which will play an essential role in the spectral theory of boundary value problems (see Chapter 7), which we shall use to solve problems that evolve in time. Let us recall first of all that, in an infinite dimensional Hilbert space, it is not true that, from every bounded sequence, we can extract a convergent subsequence (for what happens in infinite dimensions, see exercise 4.3.5).

**Theorem 4.3.21 (Rellich)** *If  $\Omega$  is an open bounded regular set of class  $C^1$ , then for every bounded sequence of  $H^1(\Omega)$  we can extract a convergent subsequence in  $L^2(\Omega)$  (we say that the canonical injection of  $H^1(\Omega)$  into  $L^2(\Omega)$  is compact).*

Theorem 4.3.21 can be false if the open set  $\Omega$  is not bounded. For example, if  $\Omega = \mathbb{R}^N$ , the canonical injection of  $H^1(\mathbb{R}^N)$  in  $L^2(\mathbb{R}^N)$  is not compact. To convince

ourselves, it is sufficient to consider the sequence  $u_n(x) = u(x + ne)$  where  $e$  is a nonzero vector and  $u$  a function of  $H^1(\mathbb{R}^N)$  (we translate  $u$  in the direction  $e$ ). It is clear that no subsequence of  $u_n$  converges in  $L^2(\mathbb{R}^N)$ .

**Remark 4.3.22** If we replace  $H^1(\Omega)$  by  $H_0^1(\Omega)$ , then not only does Rellich's Theorem 4.3.21 remain true, but further it is not necessary to assume that the open set  $\Omega$  is regular. •

The proof of theorem 4.3.21 is long and delicate (it calls on the Fourier transform, and Ascoli's compactness theorem; see, for example, [6], [34]). We shall be content with giving a simpler proof in  $N = 1$  dimension.

**Proof.** We take a space of  $N = 1$  dimension and, without loss of generality, we assume that  $\Omega = (0, 1)$ . Let  $(u_n)_{n \geq 1}$  be a bounded sequence of  $H^1(0, 1)$ , that is, there exists  $K > 0$  such that

$$\|u_n\|_{H^1(0,1)} \leq K \quad \forall n \geq 1.$$

From lemma 4.3.3, for all  $x, y \in [0, 1]$  we have

$$|u_n(x)| \leq CK \quad \text{and} \quad |u_n(x) - u_n(y)| \leq CK\sqrt{|x - y|}. \quad (4.16)$$

(We have applied the Cauchy-Schwarz inequality to (4.8).) Take a (countable) sequence  $(x_p)_{p \geq 1}$  of points of  $[0, 1]$  which is dense in this interval (for example, the points of  $\mathbb{Q} \cap [0, 1]$ ). For fixed  $p$ , the sequence  $u_n(x_p)$  is bounded in  $\mathbb{R}$  because of the first inequality of (4.16). We can therefore extract a subsequence which converges in  $\mathbb{R}$ . We first apply this procedure to the sequence  $u_n(x_1)$  and we obtain a subsequence  $u_{n_1}(x_1)$  which converges in  $\mathbb{R}$ . Then we extract from this subsequence, indexed by  $n_1$ , a new subsequence, indexed by  $n_2$ , such that  $u_{n_2}(x_2)$  converges in  $\mathbb{R}$  (but, of course, we have also  $u_{n_2}(x_1)$  which converges). By successively extracting a subsequence from the preceding, by recurrence, we construct a subsequence, indexed by  $n_p$ , such that  $u_{n_p}(x_p)$  converges, as does  $u_{n_p}(x_k)$  for  $1 \leq k \leq p$ . Obviously, the subsequences  $u_{n_p}$  are increasingly 'thin' as  $p$  becomes large. To avoid the problem that nothing remains 'in the limit', we use an argument of extraction of a diagonal sequence. In other words, we extract a last subsequence (called diagonal) from this set of subsequences, where we choose the first element of the first subsequence  $u_{n_1}$ , then the second element of the second  $u_{n_2}$ , and so on till the  $p$ th element of the  $p$ th  $u_{n_p}$ . The sequence we obtain is denoted by  $u_m$  and we see that, for all  $p \geq 1$ , the sequence  $u_m(x_p)$  converges in  $\mathbb{R}$ .

Now let  $x \in [0, 1]$  and  $\epsilon > 0$  (a small parameter). As the sequence  $(x_p)_{p \geq 1}$  is dense in  $[0, 1]$ , there exists a point  $x_p$  such that  $|x - x_p| \leq \epsilon$ . In addition, since  $u_m(x_p)$  converges in  $\mathbb{R}$  as  $m$  tends to infinity, there exists  $m_0$  such that, for all  $m, m' \geq m_0$ , we have  $|u_m(x_p) - u_{m'}(x_p)| \leq \epsilon$ . Consequently, we obtain

$$\begin{aligned} |u_m(x) - u_{m'}(x)| &\leq |u_m(x_p) - u_m(x)| + |u_m(x_p) - u_{m'}(x_p)| + |u_{m'}(x_p) - u_{m'}(x)| \\ &\leq \epsilon + 2CK\sqrt{\epsilon} \end{aligned}$$



which proves that the sequence  $u_m(x)$  is Cauchy in  $\mathbb{R}$ , and therefore converges for all  $x \in [0, 1]$ . By application of the Lebesgue dominated convergence theorem, we conclude that the sequence  $u_m$  converges in  $L^2(0, 1)$ .  $\square$

**Exercise 4.3.5** Take  $\Omega = (0, 1)$  and  $u_n(x) = \sin(2\pi nx)$ . Show that the sequence  $u_n$  is uniformly bounded in  $L^2(\Omega)$ , but that there does not exist any convergent subsequence. For this show, thanks to integration by parts, that, for every function  $\phi \in C_c^\infty(\Omega)$ , we have

$$\lim_{n \rightarrow +\infty} \int_0^1 u_n(x) \phi(x) dx = 0,$$

and we deduce a contradiction if a subsequence of  $u_n$  converges in  $L^2(\Omega)$ . Generalize this counterexample to  $H^1(\Omega)$  by considering a primitive of  $u_n$ .

### 4.3.5 The spaces $H^m(\Omega)$

We can easily generalize the definition 4.3.1 of the Sobolev space  $H^1(\Omega)$  to functions which are  $m \geq 0$  times differentiable in the weak sense. We start by giving a useful convention. Let  $\alpha = (\alpha_1, \dots, \alpha_N)$  be a **multi-index**, that is, a vector of  $N$  components which are non-negative integers  $\alpha_i \geq 0$ . We denote by  $|\alpha| = \sum_{i=1}^N \alpha_i$  and, for a function  $v$ ,

$$\partial^\alpha v(x) = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \dots \partial x_N^{\alpha_N}}(x).$$

From the definition 4.2.3 of the first weak derivative, we define by recurrence over  $m$  the weak derivative of order  $m$ : we say that a function  $v \in L^2(\Omega)$  is  $m$  times differentiable in the weak sense if all of its weak partial derivatives of order  $m - 1$  are weakly differentiable in the sense of the definition 4.2.3. We remark that, in the definition of a mixed derivative, the order of differentiation is not important, because of the Schwarz theorem  $\frac{\partial^2 v}{\partial x_i \partial x_j} = \frac{\partial^2 v}{\partial x_j \partial x_i}$ , which justifies the notation  $\partial^\alpha v$  where the order of differentiation is not given.

**Definition 4.3.23** For an integer  $m \geq 0$ , the Sobolev space  $H^m(\Omega)$  is defined by

$$H^m(\Omega) = \{v \in L^2(\Omega) \text{ such that, } \forall \alpha \text{ with } |\alpha| \leq m, \partial^\alpha v \in L^2(\Omega)\}, \quad (4.17)$$

where the partial derivative  $\partial^\alpha v$  is taken in the weak sense.

We leave the reader the task of verifying this easy result.

**Proposition 4.3.24** Equipped with the scalar product

$$\langle u, v \rangle = \int_\Omega \sum_{|\alpha| \leq m} \partial^\alpha u(x) \partial^\alpha v(x) dx \quad (4.18)$$

and with the norm  $\|u\|_{H^m(\Omega)} = \sqrt{\langle u, u \rangle}$ , the Sobolev space  $H^m(\Omega)$  is a Hilbert space.

The functions of  $H^m(\Omega)$  are not always continuous or regular (this depends on  $m$  and on the dimension  $N$ ), but if  $m$  is sufficiently large then every function of  $H^m(\Omega)$  is continuous. Let us recall that from lemma 4.3.3, in dimension  $N = 1$ , the functions of  $H^1(\Omega)$  are continuous. We have the following result which generalizes lemma 4.3.3 to higher dimensions (see lemma 4.4.9 for the simpler proof of a similar result).

**Theorem 4.3.25** *If  $\Omega$  is an open bounded regular set of class  $\mathcal{C}^1$ , and if  $m > N/2$ , then  $H^m(\Omega)$  is a subspace of the set  $C(\overline{\Omega})$  of continuous functions over  $\overline{\Omega}$ .*

**Remark 4.3.26** By repeated application of the Theorem 4.3.25 to a function and to its derivatives, we can improve this conclusion. If there exists an integer  $k \geq 0$  such that  $m - N/2 > k$ , then  $H^m(\Omega)$  is a subspace of the set  $C^k(\overline{\Omega})$  of functions  $k$  times differentiable over  $\overline{\Omega}$ . •

The ‘moral’ of theorem 4.3.25 is that the larger that  $m$  becomes, the more the functions of  $H^m(\Omega)$  are regular, that is differentiable in the usual sense (it is enough to successively apply theorem 4.3.25 to a function  $v \in H^m(\Omega)$  and to its derivatives  $\partial^\alpha v \in H^{m-|\alpha|}(\Omega)$ ).

As is the case for  $H^1(\Omega)$ , the regular functions are dense in  $H^m(\Omega)$  (if at least the open set  $\Omega$  is regular; see definition 3.2.5). The proof of the density theorem 4.3.5 generalizes very easily to  $H^m(\Omega)$ . We do not repeat it and we only state the following density result.

**Theorem 4.3.27** *If  $\Omega$  is an open bounded regular set of class  $\mathcal{C}^m$ , or if  $\Omega = \mathbb{R}_+^N$ , then  $C_c^\infty(\overline{\Omega})$  is dense in  $H^m(\Omega)$ .*

We can also obtain some trace results and Green’s formulas of higher order for the space  $H^m(\Omega)$ . For simplicity, we content ourselves with treating the case  $m = 2$  (which is the only one we use in what follows).

**Theorem 4.3.28** *Let  $\Omega$  be an open bounded regular set of class  $\mathcal{C}^1$ . We define the trace mapping  $\gamma_1$*

$$\begin{aligned} H^2(\Omega) \cap C^1(\overline{\Omega}) &\rightarrow L^2(\partial\Omega) \cap C(\overline{\partial\Omega}) \\ v &\rightarrow \gamma_1(v) = \frac{\partial v}{\partial n} \Big|_{\partial\Omega}, \end{aligned} \quad (4.19)$$

with  $\frac{\partial v}{\partial n} = \nabla u \cdot n$ . This mapping  $\gamma_1$  is extended by continuity to a continuous linear mapping of  $H^2(\Omega)$  into  $L^2(\partial\Omega)$ . In particular, there exists a constant  $C > 0$  such that, for every function  $v \in H^2(\Omega)$ , we have

$$\left\| \frac{\partial v}{\partial n} \right\|_{L^2(\partial\Omega)} \leq C \|v\|_{H^2(\Omega)}. \quad (4.20)$$

**Proof.** The existence of the trace mapping  $\gamma_1$  (and its properties) is a simple consequence of the preceding trace theorem 4.3.13 for the functions of  $H^1(\Omega)$ . In effect, if  $v \in H^2(\Omega)$ , then  $\nabla v \in H^1(\Omega)^N$  and we can therefore define the trace of  $\nabla v$  on  $\partial\Omega$  as a function of  $L^2(\partial\Omega)^N$ . As the normal is a continuous function bounded over  $\partial\Omega$ , we deduce that  $\frac{\partial v}{\partial n} \in L^2(\partial\Omega)$ .  $\square$

**Remark 4.3.29** If  $\Omega$  is an open bounded regular set of class  $C^2$ , we can improve the above trace theorem 4.3.13. We redefine the trace mapping  $\gamma_0$

$$\begin{aligned} H^2(\Omega) \cap C(\overline{\Omega}) &\rightarrow H^1(\partial\Omega) \cap C(\overline{\partial\Omega}) \\ v &\rightarrow \gamma_0(v) = v|_{\partial\Omega}, \end{aligned} \quad (4.21)$$

which is prolonged by continuity to a continuous linear mapping of  $H^2(\Omega)$  into  $H^1(\partial\Omega)$ . In other words, the trace  $\gamma_0(v)$  has tangential derivatives. If  $\Omega = \mathbb{R}_+^N$ , this result is easy enough to conceive and to prove. In the general case, it is necessary to know how to define a Sobolev space on a variety (here  $\partial\Omega$ ) which is beyond the scope of this course. Nevertheless, this is not without interest since it allows us to study models of partial differential equations which ‘live’ at the same time in a domain and on its boundary (for example, models of volume and surface diffusion, or even a model of elasticity in volume coupled with a shell model on the surface).  $\bullet$

The trace theorem 4.3.28 allows us to generalize to  $H^2(\Omega)$  a Green’s formula established for functions of class  $C^2$  in the corollary 3.2.4.

**Theorem 4.3.30** *Let  $\Omega$  be an open bounded regular set of class  $C^2$ . If  $u \in H^2(\Omega)$  and  $v \in H^1(\Omega)$ , we have*

$$\int_{\Omega} \Delta u(x) v(x) dx = - \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \frac{\partial u}{\partial n}(x) v(x) ds. \quad (4.22)$$

**Proof.** As (4.22) is true for functions of class  $C^2$  and the regular functions are dense in  $H^2(\Omega)$  and  $H^1(\Omega)$ , we use a density argument. We refer to the proof of theorem 4.3.15 for more details. The only new argument here is that we must use the continuity of the trace mapping  $\gamma_1$ , that is, the inequality (4.20).  $\square$

## 4.4 Some useful extra results

This section can be omitted in the first reading.

### 4.4.1 Proof of the density theorem 4.3.5

We start with the case  $\Omega = \mathbb{R}^N$  which is the most simple.

**Theorem 4.4.1** *The space  $C_c^\infty(\mathbb{R}^N)$  of functions of class  $C^\infty$  with compact support in  $\mathbb{R}^N$  is dense in  $H^1(\mathbb{R}^N)$ .*

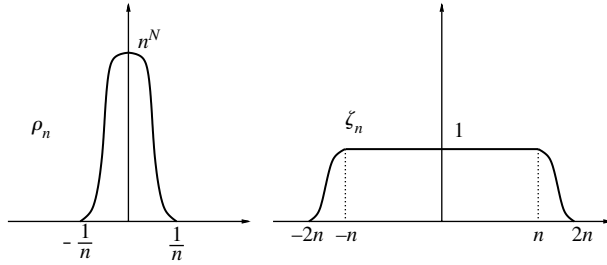


Figure 4.3. Regularization functions  $\rho_n$  (left), and truncation functions  $\zeta_n$  (right).

**Proof.** The proof is made by regularization and truncation. Take  $\rho \in C_c^\infty(B)$  (with  $B$  the unit ball) such that  $\rho \geq 0$  and  $\int_B \rho(x) dx = 1$ . We define a ‘regularizing’ sequence  $\rho_n(x) = n^N \rho(nx)$  whose support is contained in the ball of radius  $1/n$  (see Figure 4.3). We regularize, by convolution, a function  $v \in H^1(\mathbb{R}^N)$  by defining

$$v_n(x) = v \star \rho_n(x) = \int_{\mathbb{R}^N} \rho_n(x-y)v(y) dy,$$

which is of class  $C^\infty$  and such that  $\nabla v_n = (\nabla v) \star \rho_n$ . We easily verify that  $v_n$  (respectively  $\nabla v_n$ ) converges to  $v$  (respectively  $\nabla v$ ) in  $L^2(\mathbb{R}^N)$ : it is obvious when  $v$  (respectively  $\nabla v$ ) is continuous, and from the density of regular functions in  $L^2(\mathbb{R}^N)$  (see theorem 4.2.1) the result extends to every function of  $L^2(\mathbb{R}^N)$ . It only remains to truncate the sequence  $v_n$  in order to give it compact support. Let  $\zeta \in C_c^\infty(\mathbb{R}^N)$  such that  $0 \leq \zeta \leq 1$  and  $\zeta(x) = 1$  if  $|x| \leq 1$ ,  $\zeta(x) = 0$  if  $|x| \geq 2$ . We set  $\zeta_n(x) = \zeta(\frac{x}{n})$  (see Figure 4.3) and we truncate  $v_n$  by defining  $\tilde{v}_n(x) = v_n(x)\zeta_n(x)$ . We also easily verify that  $\tilde{v}_n$  (respectively  $\nabla \tilde{v}_n$ ) converges to  $v$  (respectively  $\nabla v$ ) in  $L^2(\mathbb{R}^N)$ .  $\square$

The density theorem 4.3.5 for a regular open set or for the half-space is an immediate consequence of the following result combined with the density theorem 4.4.1 in the whole space  $\mathbb{R}^N$ .

**Proposition 4.4.2** *If  $\Omega$  is an open bounded regular set of class  $C^1$ , or if  $\Omega = \mathbb{R}_+^N$ , then there exists a prolongation operator  $P$  of  $H^1(\Omega)$  in  $H^1(\mathbb{R}^N)$  which is a continuous linear mapping such that, for all  $v \in H^1(\Omega)$ ,*

- (1)  $Pv|_\Omega = v$ ,
- (2)  $\|Pv\|_{L^2(\mathbb{R}^N)} \leq C\|v\|_{L^2(\Omega)}$ ,
- (3)  $\|Pv\|_{H^1(\mathbb{R}^N)} \leq C\|v\|_{H^1(\Omega)}$ ,

where the constant  $C > 0$  depends only on  $\Omega$ .

**Proof.** First, we prove the result for  $\Omega = \mathbb{R}_+^N$ . We denote  $x = (x', x_N)$  with  $x' = (x_1, \dots, x_{N-1})$ . Take  $v \in H^1(\mathbb{R}_+^N)$ . We define

$$Pv(x) = \begin{cases} v(x', x_N) & \text{if } x_N > 0 \\ v(x', -x_N) & \text{if } x_N < 0. \end{cases}$$

We verify then that, for  $1 \leq i \leq N-1$ ,

$$\frac{\partial P v}{\partial x_i}(x) = \begin{cases} \frac{\partial v}{\partial x_i}(x', x_N) & \text{if } x_N > 0 \\ \frac{\partial v}{\partial x_i}(x', -x_N) & \text{if } x_N < 0, \end{cases}$$

and that

$$\frac{\partial P v}{\partial x_N}(x) = \begin{cases} \frac{\partial v}{\partial x_N}(x', x_N) & \text{if } x_N > 0 \\ -\frac{\partial v}{\partial x_N}(x', -x_N) & \text{if } x_N < 0. \end{cases}$$

These equations are obvious if  $v$  is a regular function, but need justification when  $v$  is only weakly differentiable. Here, we shall not detail the (easy) arguments which justify these equations (in particular we must use the symmetry under reflection of  $Pv$ ; see [34] for the details). We therefore deduce the desired properties for the prolongation operator  $P$  with the constant  $C = \sqrt{2}$  (in the case  $\Omega = \mathbb{R}_+^N$ ).

If  $\Omega$  is an open bounded regular set of class  $C^1$ , we use a ‘local coordinate’ argument to return to the case  $\Omega = \mathbb{R}_+^N$ . By using the notation of the definition 3.2.5 of a regular open set, there exists a finite covering of  $\Omega$  by open sets  $(\omega_i)_{0 \leq i \leq I}$ . We then introduce a ‘partition of unity’ associated with this covering, that is, functions  $(\theta_i)_{0 \leq i \leq I}$  of  $C_c^\infty(\mathbb{R}^N)$  such that

$$\theta_i \in C_c^\infty(\omega_i), \quad 0 \leq \theta_i(x) \leq 1, \quad \sum_{i=0}^I \theta_i(x) = 1 \text{ in } \overline{\Omega}.$$

(The existence of such a partition of unity is classical: see Theorem 3.2.9 in [4].) We shall define  $Pv$  in the form

$$Pv = \sum_{i=0}^I P_i(\theta_i v),$$

where each operator  $P_i$  is defined locally in  $\omega_i$ . As  $\theta_0 v$  has compact support in  $\Omega$ , we define  $P_0(\theta_0 v)$  as the extension of  $\theta_0 v$  by zero outside of  $\Omega$ . For every  $i \in \{1, \dots, I\}$ , denoting by  $\phi_i$  the mapping which transforms  $\omega_i$  into a reference domain  $Q$  (see definition 3.2.5 and Figure 4.4), we set

$$w_i = (\theta_i v) \circ (\phi_i^{-1}|_{Q^+}) \quad \text{with } Q^+ = Q \cap \mathbb{R}_+^N.$$

This function  $w_i$  belongs to  $H^1(Q^+)$  and is zero in a neighbourhood of  $\partial Q^+ \cap \mathbb{R}_+^N$ . If we extend by 0 in  $\mathbb{R}_+^N \setminus Q^+$ , we obtain a function  $\tilde{w}_i \in H^1(\mathbb{R}_+^N)$ . We can then extend  $\tilde{w}_i$  by reflection to obtain a function  $P\tilde{w}_i \in H^1(\mathbb{R}^N)$  (we use the prolongation operator  $P$  that we have just constructed for  $\mathbb{R}_+^N$ ). We return to  $\omega_i$  and we set

$$P_i(\theta_i v) = (P\tilde{w}_i) \circ \phi_i.$$

By the  $C^1$  regularity of  $\phi_i$  and of its inverse, we obtain the desired properties for  $P_i$  and therefore  $P$ .  $\square$

**Remark 4.4.3** The ‘local coordinate’ argument used above is very classical and is used in many proofs. As it is technical, we sometimes use it without giving more details.  $\bullet$

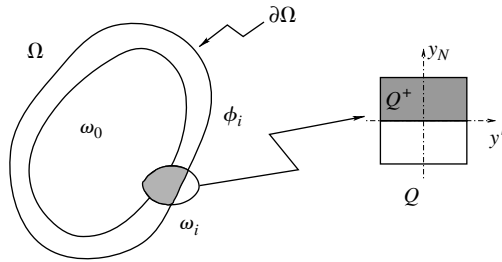


Figure 4.4. Local coordinates of a regular open set.

#### 4.4.2 The space $H(\text{div})$

We introduce another space, intermediate between  $L^2(\Omega)$  and  $H^1(\Omega)$ , for vector valued functions. This space is very useful in certain applications (see, for example, remark 5.2.15).

**Definition 4.4.4** *The space  $H(\text{div})$  is defined by*

$$H(\text{div}) = \{ \sigma \in L^2(\Omega)^N \text{ such that } \text{div} \sigma \in L^2(\Omega) \}, \quad (4.23)$$

where  $\text{div} \sigma$  is the weak divergence of  $\sigma$  in the sense of the definition 4.2.6.

We easily verify that it is a Hilbert space (the proof is left to the reader as an exercise).

**Proposition 4.4.5** *Equipped with the scalar product*

$$\langle \sigma, \tau \rangle = \int_{\Omega} (\sigma(x) \cdot \tau(x) + \text{div} \sigma(x) \text{div} \tau(x)) dx \quad (4.24)$$

and with the norm  $\|\sigma\|_{H(\text{div})} = \sqrt{\langle \sigma, \sigma \rangle}$ , the space  $H(\text{div})$  is a Hilbert space.

As for the Sobolev spaces, we can prove a density result for regular functions (we omit the proof which is similar to that of theorem 4.3.5).

**Theorem 4.4.6** *If  $\Omega$  is an open bounded regular set of class  $C^1$ , or if  $\Omega = \mathbb{R}_+^N$ , then  $C_c^\infty(\overline{\Omega})^N$  is dense in  $H(\text{div})$ .*

One of the interests in the space  $H(\text{div})$  is that it allows us to prove a trace theorem and a Green's formula with even less regularity than in the Sobolev space  $H^1(\Omega)$ . In effect, if  $\sigma$  belongs to  $H(\text{div})$ , we only 'control' a single combination of its partial derivatives (and not all as in  $H^1(\Omega)$ ), but we can nevertheless give a meaning to the **normal trace**  $\sigma \cdot n$  on  $\partial\Omega$ .

We start by recalling that  $\gamma_0$  denotes the trace mapping of  $H^1(\Omega)$  into  $L^2(\partial\Omega)$  (see the trace theorem 4.3.13) and that  $\text{Im}(\gamma_0) = H^{1/2}(\partial\Omega)$  which is a dense subspace in  $L^2(\partial\Omega)$  (see remark 4.3.17). We can equip  $H^{1/2}(\partial\Omega)$  with the following norm

$$\|v\|_{H^{1/2}(\partial\Omega)} = \inf \{ \|\phi\|_{H^1(\Omega)} \text{ such that } \gamma_0(\phi) = v \}$$

which makes it a Banach space (and even a Hilbert space). We then define  $H^{-1/2}(\partial\Omega)$  as the dual of  $H^{1/2}(\partial\Omega)$ .

**Theorem 4.4.7 (divergence formula)** *Let  $\Omega$  be an open bounded regular set of class  $\mathcal{C}^1$ . We define the ‘normal trace’ mapping  $\gamma_n$*

$$\begin{aligned} H(\operatorname{div}) \cap C(\overline{\Omega}) &\rightarrow H^{-1/2}(\partial\Omega) \cap C(\overline{\partial\Omega}) \\ \sigma = (\sigma_i)_{1 \leq i \leq N} &\rightarrow \gamma_n(\sigma) = (\sigma \cdot n)|_{\partial\Omega} \end{aligned}$$

where  $n = (n_i)_{1 \leq i \leq N}$  is the outward unit normal to  $\partial\Omega$ . This mapping  $\gamma_n$  is extended by continuity to a continuous linear mapping from  $H(\operatorname{div})$  into  $H^{-1/2}(\partial\Omega)$ . Further, if  $\sigma \in H(\operatorname{div})$  and  $\phi \in H^1(\Omega)$ , we have

$$\int_{\Omega} \operatorname{div} \sigma \phi \, dx + \int_{\Omega} \sigma \cdot \nabla \phi \, dx = \langle \sigma \cdot n, \gamma_0(\phi) \rangle_{H^{-1/2}, H^{1/2}(\partial\Omega)}. \quad (4.25)$$

**Proof.** If  $\Omega$  is a regular open set of class  $\mathcal{C}^1$ , exercise 3.2.1 gives the following integration by parts formula, called the divergence formula,

$$\int_{\Omega} \operatorname{div} \sigma \phi \, dx + \int_{\Omega} \sigma \cdot \nabla \phi \, dx = \int_{\partial\Omega} \sigma \cdot n \phi \, ds, \quad (4.26)$$

for regular functions  $\sigma$  and  $\phi$ . We remark happily that the ‘unpleasant’ term on the right in (4.25) is none other than the usual ‘pleasant’ boundary integral in (4.26). We can see easily that the two terms on the left of (4.26) have a meaning for  $\phi \in H^1(\Omega)$  and  $\sigma \in H(\operatorname{div})$ . Then, by the density of regular functions in  $H^1(\Omega)$  and  $H(\operatorname{div})$ , the terms on the left of (4.26) are extended by continuity and the term on the right appears as a continuous linear form over the image of the trace mapping  $\operatorname{Im}(\gamma_0)$ , denoted  $H^{1/2}(\partial\Omega)$ . Such a linear form is written exactly as in the formula (4.25) and the normal trace  $\gamma_n(\sigma)$  is therefore well-defined as an element of  $H^{-1/2}(\partial\Omega)$ .  $\square$

We can, of course, also define the equivalent of  $H_0^1(\Omega)$  for the space  $H(\operatorname{div})$ . We define the subspace  $H_0(\operatorname{div})$  of  $H(\operatorname{div})$  as the closure of  $C_c^\infty(\Omega)$  in  $H(\operatorname{div})$ . This is again a Hilbert space which is interpreted (if the open set is regular) as the subspace of functions of  $H(\operatorname{div})$  whose normal trace is zero.

### 4.4.3 The spaces $W^{m,p}(\Omega)$

More generally, we can define spaces  $W^{m,p}(\Omega)$  for an integer  $m \geq 0$  and for a real  $1 \leq p \leq +\infty$ . These spaces are constructed on the Banach space  $L^p(\Omega)$  (see (4.1) and (4.2)). As we have said in remark 4.2.8, the idea of the weak derivative extends to  $L^p(\Omega)$ . We can therefore give the following definition.

**Definition 4.4.8** *For every integer  $m \geq 0$ , the Sobolev space  $W^{m,p}(\Omega)$  is defined by*

$$W^{m,p}(\Omega) = \{v \in L^p(\Omega) \text{ such that, } \forall \alpha \text{ with } |\alpha| \leq m, \partial^\alpha v \in L^p(\Omega)\}, \quad (4.27)$$

where the partial derivative  $\partial^\alpha v$  is taken in the weak sense.

Equipped with the norm

$$\|u\|_{W^{m,p}(\Omega)} = \left( \sum_{|\alpha| \leq m} \|\partial^\alpha u\|^p \right)^{1/p}$$

we verify that  $W^{m,p}(\Omega)$  is a Banach space. These spaces are particularly important for nonlinear problems (that we do not consider here, see, for example, [18] [27]), but also for linear problems because of the celebrated **Sobolev inequalities**. We state them without proof. If  $\Omega$  is a regular open set, or if  $\Omega = \mathbb{R}^N$  or  $\Omega = \mathbb{R}_+^N$ , then

$$\begin{cases} \text{if } p < N & W^{1,p}(\Omega) \subset L^q(\Omega) \ \forall q \in [1, p^*] \text{ with } 1/p^* = 1/p - 1/N \\ \text{if } p = N & W^{1,p}(\Omega) \subset L^q(\Omega) \ \forall q \in [1, +\infty[ \\ \text{if } p > N & W^{1,p}(\Omega) \subset C(\overline{\Omega}), \end{cases} \quad (4.28)$$

with continuous injection, that is,  $W^{1,p}(\Omega) \subset E$  which means that there exists a constant  $C$  such that, for all  $u \in W^{1,p}(\Omega)$ ,

$$\|u\|_E \leq C \|u\|_{W^{1,p}(\Omega)}.$$

The particular case  $p = 1$  and  $m = N$  is remarkable as we can very simply prove a Sobolev type inequality.

**Lemma 4.4.9** *The space  $W^{N,1}(\mathbb{R}^N)$  is injected continuously into the space of continuously bounded functions over  $\mathbb{R}^N$ , denoted  $C_b(\mathbb{R}^N)$ , and for all  $u \in W^{N,1}(\mathbb{R}^N)$  we have*

$$\|u\|_{L^\infty(\mathbb{R}^N)} \leq \|u\|_{W^{N,1}(\mathbb{R}^N)}. \quad (4.29)$$

**Proof.** Let  $u \in C_c^\infty(\mathbb{R}^N)$ . For  $x = (x_1, \dots, x_N)$ , we have

$$u(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} \frac{\partial^N u}{\partial x_1 \cdots \partial x_N}(y) dy_1 \cdots dy_N,$$

from which we deduce

$$\|u\|_{L^\infty(\mathbb{R}^N)} \leq \left\| \frac{\partial^N u}{\partial x_1 \cdots \partial x_N} \right\|_{L^1(\mathbb{R}^N)} \leq \|u\|_{W^{N,1}(\mathbb{R}^N)}.$$

Now  $C_c^\infty(\mathbb{R}^N)$  is dense in  $W^{N,1}(\mathbb{R}^N)$  (this is proved like the density theorem 4.4.1 for  $H^1(\mathbb{R}^N)$ ). Therefore, by density we obtain the inequality (4.29) for all  $u \in W^{N,1}(\mathbb{R}^N)$ . In addition, the closure of  $C_c^\infty(\mathbb{R}^N)$  for the norm of  $L^\infty(\mathbb{R}^N)$  is exactly  $C_b(\mathbb{R}^N)$  (in fact, these two spaces have the same norm). Therefore, (4.29) implies that the functions of  $W^{N,1}(\mathbb{R}^N)$  are continuous and bounded.  $\square$

#### 4.4.4 Duality

Recall that the dual  $V'$  of a Hilbert space  $V$  is the set of continuous linear forms over  $V$ . By application of the Riesz representation theorem 12.1.18, the dual of  $L^2(\Omega)$  is identical to  $L^2(\Omega)$  itself. We can also define the dual of a Sobolev space. In fact, the dual of  $H_0^1(\Omega)$  plays a particular role in what follows.



**Definition 4.4.10** *The dual of the Sobolev space  $H_0^1(\Omega)$  is called  $H^{-1}(\Omega)$ . We denote by  $\langle L, \phi \rangle_{H^{-1}, H_0^1(\Omega)} = L(\phi)$  the duality pairing between  $H_0^1(\Omega)$  and its dual for all continuous linear forms  $L \in H^{-1}(\Omega)$  and every function  $\phi \in H_0^1(\Omega)$ .*

We can characterize this dual  $H^{-1}(\Omega)$ . We have the following result (see [6]).

**Proposition 4.4.11** *The space  $H^{-1}(\Omega)$  is characterized by*

$$H^{-1}(\Omega) = \left\{ f = v_0 + \sum_{i=1}^N \frac{\partial v_i}{\partial x_i} \quad \text{with } v_0, v_1, \dots, v_N \in L^2(\Omega) \right\}.$$

*In other words, every continuous linear form over  $H_0^1(\Omega)$ , denoted by  $L \in H^{-1}(\Omega)$ , is written for all  $\phi \in H_0^1(\Omega)$*

$$L(\phi) = \int_{\Omega} \left( v_0 \phi - \sum_{i=1}^N v_i \frac{\partial \phi}{\partial x_i} \right) dx$$

*with  $v_0, v_1, \dots, v_N \in L^2(\Omega)$ .*

Thanks to the space  $H^{-1}(\Omega)$  we can define a new idea of differentiation for the functions of  $L^2(\Omega)$  (more weak than the weak derivative of the definition 4.2.3). Faced with this influx of notions of differentiation, we reassure the uneasy reader by saying that they are all versions of differentiation in the sense of distributions (it is precisely one of the purposes of the theory of distributions to unify these various types of differentiation).

**Lemma 4.4.12** *Take  $v \in L^2(\Omega)$ . For  $1 \leq i \leq N$ , we can define a continuous linear form  $\frac{\partial v}{\partial x_i}$  in  $H^{-1}(\Omega)$  by the formula*

$$\left\langle \frac{\partial v}{\partial x_i}, \phi \right\rangle_{H^{-1}, H_0^1(\Omega)} = - \int_{\Omega} v \frac{\partial \phi}{\partial x_i} dx \quad \forall \phi \in H_0^1(\Omega), \quad (4.30)$$

*which satisfies*

$$\left\| \frac{\partial v}{\partial x_i} \right\|_{H^{-1}(\Omega)} \leq \|v\|_{L^2(\Omega)}.$$

*If  $v \in H^1(\Omega)$ , then the continuous linear form  $\frac{\partial v}{\partial x_i}$  coincides with the weak derivative in  $L^2(\Omega)$  of  $v$ .*

**Proof.** We verify easily that the right-hand side of (4.30) is a continuous linear form over  $H_0^1(\Omega)$ . Consequently, there exists an element  $L_i \in H^{-1}(\Omega)$  such that

$$L_i(\phi) = - \int_{\Omega} v \frac{\partial \phi}{\partial x_i} dx.$$

We also verify easily that the mapping  $v \rightarrow L_i$  is continuous and linear from  $L^2(\Omega)$  into  $H^{-1}(\Omega)$ , and that it prolongs the usual differentiation of the regular functions  $v$  (or the weak differentiation for  $v \in H^1(\Omega)$ ). Consequently, we can extend, by continuity, differentiation to every function of  $L^2(\Omega)$  and denote  $L_i = \frac{\partial v}{\partial x_i}$ .  $\square$

**Remark 4.4.13** Thanks to the Riesz representation theorem we know that we can identify the dual of a Hilbert space with itself. However, in practice we never identify  $H^{-1}(\Omega)$  and  $H_0^1(\Omega)$ . In effect, we have defined  $H_0^1(\Omega)$  as a strict (but dense) subspace of  $L^2(\Omega)$ . Now we have already decided to identify  $L^2(\Omega)$  (equipped with the usual scalar product) and its dual (it is only a convention but it is universal), therefore we cannot further identify  $H^{-1}(\Omega)$  and  $H_0^1(\Omega)$  (with another scalar product). The correct (and usual) situation is therefore

$$H_0^1(\Omega) \subset L^2(\Omega) \equiv \left(L^2(\Omega)\right)' \subset H^{-1}(\Omega),$$

where the inclusions are strict. •

## 4.5 Link with distributions

All of the ideas introduced in this chapter have a link with the theory of distributions (see [4], [38]). Strictly speaking, we do not need the theory of distributions to define Sobolev spaces and solve boundary value problems (historically, distributions, as a mathematical theory, appeared after Sobolev spaces and the variational approach for solving partial differential equations). However, the theory of distributions is a unifying framework for all these spaces, and those readers who understand this theory will not fail to ask what are the links between this theory and the one we have described. As for the readers unfamiliar with this theory, they can legitimately wonder about the ideas at the base of the theory of distributions. It is to satisfy this double curiosity that we have written this section which is a (very) brief and sketchy summary of the theory of distributions.

Let  $\Omega$  be an open set of  $\mathbb{R}^N$ . We denote by  $C_c^\infty(\Omega)$  (or  $\mathcal{D}(\Omega)$ ) the space of functions of class  $C^\infty$  with compact support in  $\Omega$ . We equip  $C_c^\infty(\Omega)$  with a ‘pseudo-topology’, that is, we define a notion of convergence in  $C_c^\infty(\Omega)$ . We say that a sequence  $(\phi_n)_{n \geq 1}$  of  $C_c^\infty(\Omega)$  converges to  $\phi \in C_c^\infty(\Omega)$  if

- (1) the support of  $\phi_n$  remains in a compact set  $K$  of  $\Omega$ ,
- (2) for all multi-indices  $\alpha$ ,  $\partial^\alpha \phi_n$  converges uniformly in  $K$  to  $\partial^\alpha \phi$ .

**The space of distributions  $\mathcal{D}'(\Omega)$**  is the ‘dual’ of  $\mathcal{D}(\Omega)$ , that is, the space of ‘continuous’ linear forms over  $\mathcal{D}(\Omega)$ . The quotation marks are used as there is not a norm defined over  $\mathcal{D}(\Omega)$ , but simply a notion of convergence. Nevertheless, we can precisely define  $\mathcal{D}'(\Omega)$  whose elements are called **distributions**.

**Definition 4.5.1** A distribution  $T \in \mathcal{D}'(\Omega)$  is a linear form over  $\mathcal{D}(\Omega)$  which satisfies

$$\lim_{n \rightarrow +\infty} T(\phi_n) = T(\phi)$$

for every sequence  $(\phi_n)_{n \geq 1}$  of  $C_c^\infty(\Omega)$  which converges to  $\phi \in C_c^\infty(\Omega)$  in the sense defined above.

We denote by  $\langle T, \phi \rangle = T(\phi)$  the duality pairing between a distribution  $T \in \mathcal{D}'(\Omega)$  and a function  $\phi \in \mathcal{D}(\Omega)$ : this duality pairing ‘generalizes’ the usual integral  $\int_\Omega T \phi dx$ . In effect, we verify that if  $f$  is a function (which is locally integrable in  $\Omega$ ), then we can define a distribution  $T_f$  by

$$\langle T_f, \phi \rangle = \int_\Omega f \phi dx.$$

Consequently, we identify the function and the associated distribution  $T_f \equiv f$ .

We can also give  $\mathcal{D}'(\Omega)$  an idea of convergence: we say that a sequence  $T_n \in \mathcal{D}'(\Omega)$  **converges in the sense of distributions** to  $T \in \mathcal{D}'(\Omega)$  if, for all  $\phi \in \mathcal{D}(\Omega)$ ,

$$\lim_{n \rightarrow +\infty} \langle T_n, \phi \rangle = \langle T, \phi \rangle.$$

This convergence in the sense of distributions is an extremely weak (or not very demanding) convergence as it corresponds to an integral, or ‘average’ convergence.

Let us now define **differentiation in the sense of distributions**: if  $T \in \mathcal{D}'(\Omega)$ , we define  $\frac{\partial T}{\partial x_i} \in \mathcal{D}'(\Omega)$  by

$$\left\langle \frac{\partial T}{\partial x_i}, \phi \right\rangle = - \left\langle T, \frac{\partial \phi}{\partial x_i} \right\rangle \quad \forall \phi \in \mathcal{D}(\Omega).$$

We verify that, effectively, the derivative  $\partial T / \partial x_i$  is a distribution, that is, distributions are infinitely differentiable! This is one of the most important properties of distributions. We verify also that if  $f$  is a function which is differentiable in the classical sense, then its derivative in the sense of distributions coincides with its usual derivative.

Of course, we recognize in weak differentiation in the sense of definition 4.2.3 a particular case of differentiation in the sense of distributions. Further, all the spaces  $L^p(\Omega)$  or the Sobolev spaces  $H^m(\Omega)$  are subspaces of the space of distributions  $\mathcal{D}'(\Omega)$ . In particular, we verify that convergence in these spaces implies convergence in the sense of distributions (but the converse is false). Finally, the equations in the variational formulations (which we have interpreted as equations almost everywhere) again imply, more simply and more generally, equations in the sense of distributions.

Lemma 4.2.4 (weak differentiation)	$u \in L^2(\Omega)$ is differentiable in the weak sense if, $\forall i$ , $\left  \int_{\Omega} u \frac{\partial \phi}{\partial x_i} dx \right  \leq C \ \phi\ _{L^2(\Omega)} \quad \forall \phi \in C_c^\infty(\Omega)$
Proposition 4.3.2	$H^1(\Omega)$ is a Hilbert space for the scalar product $\langle u, v \rangle = \int_{\Omega} (\nabla u \cdot \nabla v + uv) dx$
Theorem 4.3.5 (density theorem)	$C_c^\infty(\overline{\Omega})$ is dense in $H^1(\Omega)$
Proposition 4.3.10 (Poincaré inequality)	$\forall u \in H_0^1(\Omega)$ ( $\Omega$ bounded) $\ u\ _{L^2(\Omega)} \leq C \ \nabla u\ _{L^2(\Omega)}$
Theorem 4.3.13 (trace theorem)	$u \rightarrow u _{\partial\Omega}$ is a continuous mapping of $H^1(\Omega)$ into $L^2(\partial\Omega)$
Theorem 4.3.15 (Green's formula)	$\forall u, v \in H^1(\Omega)$ $\int_{\Omega} u \frac{\partial v}{\partial x_i} dx = - \int_{\Omega} v \frac{\partial u}{\partial x_i} dx + \int_{\partial\Omega} uv n_i ds$
Corollary 4.3.16 (characterization of $H_0^1(\Omega)$ )	$H_0^1(\Omega)$ is the subspace of functions of $H^1(\Omega)$ which are zero on $\partial\Omega$
Theorem 4.3.21 (Rellich theorem)	The injection of $H^1(\Omega)$ in $L^2(\Omega)$ is compact ( $\Omega$ bounded and regular)
Theorem 4.3.30 (Green's formula)	$\forall u \in H^2(\Omega), v \in H^1(\Omega)$ $\int_{\Omega} v \Delta u dx = - \int_{\Omega} \nabla u \cdot \nabla v dx + \int_{\partial\Omega} \frac{\partial u}{\partial n} v ds$

Table 4.1. Principal results on Sobolev spaces which must be known

*This page intentionally left blank*

# 5 Mathematical study of elliptic problems

---

## 5.1 Introduction

In this chapter we shall finish the mathematical analysis of elliptic partial differential equations (PDEs) which was started in chapter 3. To show that boundary value problems are well-posed for these elliptic PDEs, that is, they have a solution, which is unique, and depends continuously on the data, we follow the **variational approach** presented in Chapter 3 and we use the **Sobolev spaces** introduced in Chapter 4.

The plan of this chapter is the following. In Section 5.2 we explain in detail the functioning of the variational approach for the Laplacian with various types of boundary condition. We show **existence and uniqueness results for the solutions**. We also show that these solutions **minimize an energy** and satisfy a number of **qualitative properties** which are very natural and important from the point of view of applications (maximum principle, regularity). Section 5.3 follows the same programme but for other, more complicated, models like that of **linear elasticity** or of the **Stokes equations**. If the existence and uniqueness theory is very simple as in the preceding case, it is not the same for all of the qualitative properties

## 5.2 Study of the Laplacian

### 5.2.1 Dirichlet boundary conditions

We consider the following boundary value problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (5.1)$$

where  $\Omega$  is an open bounded set of the space  $\mathbb{R}^N$ , and  $f$  is the right-hand side of the equation which belongs to the space  $L^2(\Omega)$ . The variational approach to study (5.1) is composed of three stages which we detail.

### Stage 1: Establishment of a variational formulation

In the first stage we must propose a variational formulation of the boundary value problem (5.1), that is, we must find a bilinear form  $a(\cdot, \cdot)$ , a linear form  $L(\cdot)$ , and a Hilbert space  $V$  such that (5.1) is equivalent to:

$$\text{Find } u \in V \text{ such that } a(u, v) = L(v) \text{ for all } v \in V. \quad (5.2)$$

The aim of this first stage is only to find the variational formulation (5.2); we shall verify the precise equivalence with (5.1) later in the course of the third stage.

To find the variational formulation we multiply equation (5.1) by a regular test function  $v$  and we integrate by parts. This calculation is mainly formal in the sense that we assume the existence and regularity of the solution  $u$  so that all the calculations carried out are allowable. With the help of Green's formula (4.22) (see also (3.7)) we find

$$\int_{\Omega} f v \, dx = - \int_{\Omega} \Delta u v \, dx = \int_{\Omega} \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, ds. \quad (5.3)$$

As  $u$  must satisfy a Dirichlet boundary condition,  $u = 0$  on  $\partial\Omega$ , we choose a Hilbert space  $V$  such that every function  $v \in V$  also satisfies  $v = 0$  on  $\partial\Omega$ . In this case, the equation (5.3) becomes

$$\int_{\Omega} \nabla u(x) \cdot \nabla v(x) \, dx = \int_{\Omega} f(x) v(x) \, dx. \quad (5.4)$$

In order that the term on the left of (5.4) has a meaning it is sufficient that  $\nabla u$  and  $\nabla v$  belong to  $L^2(\Omega)$  (component by component), and in order that the term on the right of (5.4) also has a meaning it is sufficient that  $v$  belongs to  $L^2(\Omega)$  (we have assumed that  $f \in L^2(\Omega)$ ). Consequently, a reasonable choice for the Hilbert space is  $V = H_0^1(\Omega)$ , the subspace of  $H^1(\Omega)$  whose elements are zero on the boundary  $\partial\Omega$ .

To conclude, the proposed variational formulation for (5.1) is:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \, \forall v \in H_0^1(\Omega). \quad (5.5)$$

Obviously, we have made some choices to arrive at (5.5); other choices would have led us to other possible variational formulations. The justification of (5.5) is therefore carried out *a posteriori*. The next stage consists of verifying that (5.5) has a unique solution, then the third stage that the solution of (5.5) is also a solution of the boundary value problem (5.1) (in a sense to be made precise).

## Stage 2: Solution of the variational formulation

In this second stage we verify that the variational formulation (5.5) has a unique solution. For this we use the Lax–Milgram theorem 3.3.1 whose hypotheses we check with the notation

$$a(u, v) = \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx \text{ and } L(v) = \int_{\Omega} f(x)v(x) dx.$$

We easily see by using the Cauchy–Schwarz inequality that  $a$  is a continuous bilinear form over  $H_0^1(\Omega)$  and that  $L$  is a continuous linear form over  $H_0^1(\Omega)$ . Further, from the Poincaré inequality (see corollary 4.3.12; we here use the bounded character of the open set  $\Omega$ ), the bilinear form  $a$  is coercive, that is, there exists  $\nu > 0$  such that

$$a(v, v) = \int_{\Omega} |\nabla v(x)|^2 dx \geq \nu \|v\|_{H_0^1(\Omega)}^2 \quad \forall v \in H_0^1(\Omega).$$

As  $H_0^1(\Omega)$  is a Hilbert space (see proposition 4.3.9), all the hypotheses of the Lax–Milgram theorem 3.3.1 are satisfied and we can therefore conclude that there exists a unique solution  $u \in H_0^1(\Omega)$  of the variational formulation (5.5).

**Remark 5.2.1** We see later in Chapter 9 that, in the present case, since the bilinear form  $a$  is symmetric, there exists an argument other than the Lax–Milgram theorem to reach the conclusion. In effect, the solution of the variational formulation is, in this case, the unique minimum of the energy defined by

$$J(v) = \frac{1}{2}a(v, v) - L(v) \quad \forall v \in H_0^1(\Omega)$$

(see proposition 5.2.7). Consequently, if we prove that  $J$  has a unique minimum, we have therefore obtained the solution of the variational formulation. •

## Stage 3: Equivalence with the equation

The third stage (the last and most delicate) consists of verifying that if we solve the variational formulation (5.5) we have solved the boundary value problem (5.1), and making precise in what sense the solution of (5.5) is also a solution of (5.1). In other words, it is a question of interpreting the variational formulation and of returning to the equation. For this we use the same integrations by parts which led to the variational formulation, but in the opposite sense, and justify them carefully.

This justification is very easy if we assume that the solution  $u$  of the variational formulation (5.5) is regular (more precisely if  $u \in H^2(\Omega)$ ) and that the open set  $\Omega$  is also regular, which we do initially. In effect, it is enough to use Green’s formula (4.22) which yields, for  $v \in H_0^1(\Omega)$ ,

$$\int_{\Omega} \nabla u \cdot \nabla v dx = - \int_{\Omega} v \Delta u dx$$

as  $v = 0$  on the boundary  $\partial\Omega$ . We then deduce

$$\int_{\Omega} (\Delta u + f) v dx = 0 \quad \forall v \in C_c^\infty(\Omega),$$



which implies, from corollary 4.2.2, that  $-\Delta u = f$  in  $L^2(\Omega)$ , and we have the equation

$$-\Delta u = f \text{ almost everywhere in } \Omega. \quad (5.6)$$

Further, if  $\Omega$  is a regular open bounded set of class  $\mathcal{C}^1$ , then the trace theorem 4.3.13 (or more precisely its corollary 4.3.16) confirms that every function of  $H_0^1(\Omega)$  has a trace on  $\partial\Omega$  which is zero in  $L^2(\Omega)$ . We deduce, in particular, that

$$u = 0 \text{ almost everywhere on } \partial\Omega. \quad (5.7)$$

We have therefore recovered the equation and the boundary conditions of (5.1).

If we no longer assume that the solution  $u$  of (5.5) and the open set  $\Omega$  are regular, we must work harder (we can no longer use Green's formula (4.22) which needs  $u \in H^2(\Omega)$ ). We denote  $\sigma = \nabla u$  which is a vector valued function in  $L^2(\Omega)^N$ . By the Cauchy-Schwarz inequality, we deduce from the variational formulation (5.5) that, for all  $v \in H_0^1(\Omega)$ ,

$$\left| \int_{\Omega} \sigma \cdot \nabla v \, dx \right| = \left| \int_{\Omega} f v \, dx \right| \leq C \|v\|_{L^2(\Omega)}. \quad (5.8)$$

Since  $C_c^\infty(\Omega) \subset H_0^1(\Omega)$ , (5.8) is none other than the criterion for existence of a weak divergence of  $\sigma$  in  $L^2(\Omega)$  (see definition 4.2.6 and lemma 4.2.7) which satisfies, for all  $v \in C_c^\infty(\Omega)$ ,

$$\int_{\Omega} \sigma \cdot \nabla v \, dx = - \int_{\Omega} \operatorname{div} \sigma v \, dx.$$

We therefore deduce that

$$\int_{\Omega} (\operatorname{div} \sigma + f) v \, dx = 0 \quad \forall v \in C_c^\infty(\Omega),$$

which implies, from corollary 4.2.2, that  $-\operatorname{div} \sigma = f$  in  $L^2(\Omega)$ . Consequently,  $\operatorname{div} \sigma = \Delta u$  belongs to  $L^2(\Omega)$  (recall that  $\operatorname{div} \nabla = \Delta$ ), and we recover the equation (5.6). We recover the boundary conditions (5.7) as before if the open set  $\Omega$  is regular of class  $\mathcal{C}^1$ . If  $\Omega$  is not regular, then we cannot use the trace theorem 4.3.13 to obtain (5.7). Nevertheless, the simple fact of belonging to  $H_0^1(\Omega)$  is a generalization of the Dirichlet boundary condition for a nonregular open set, and we continue to write **formally** that  $u = 0$  on  $\partial\Omega$ .

To conclude we have proved the following result.

**Theorem 5.2.2** *Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$ . Take  $f \in L^2(\Omega)$ . There exists a unique solution  $u \in H_0^1(\Omega)$  of the variational formulation (5.5). Further,  $u$  satisfies*

$$-\Delta u = f \text{ almost everywhere in } \Omega \quad \text{and} \quad u \in H_0^1(\Omega). \quad (5.9)$$

*If we assume further that  $\Omega$  is regular of class  $\mathcal{C}^1$ , then  $u$  is a solution of the boundary value problem (5.1) in the sense that*

$$-\Delta u = f \text{ almost everywhere in } \Omega, \quad u = 0 \text{ almost everywhere on } \partial\Omega.$$

We call the solution  $u \in H_0^1(\Omega)$  of the variational formulation (5.5) the **variational solution** of the boundary value problem (5.1). By a convenient abuse of language, **we shall say that the unique solution  $u \in H_0^1(\Omega)$  of the variational formulation (5.5) is the unique solution of the boundary value problem (5.1)**. This use is justified by theorem 5.2.2.

The variational solution of (5.1) only satisfies the equation and the boundary conditions *a priori* in a ‘weak’ sense, that is, almost everywhere (or even worse for the boundary condition if the open set is not regular). We then talk of a **weak solution** as opposed to the strong solutions that we could have hoped to obtain in the classical formulation of (5.1) (see section 3.1.2). Likewise, we sometimes call the variational formulation the **weak formulation** of the equation.

**Remark 5.2.3** In fact, the weak solution may be a strong solution if the right-hand side  $f$  is more regular. In other words, the equation and the boundary conditions of (5.1) may be satisfied in a classical sense, that is, for all  $x \in \Omega$ , and all  $x \in \partial\Omega$ , respectively. This is what we call a regularity result for the solution (see later corollary 5.2.27). •

**Remark 5.2.4** We must understand the exact meaning of the expression  $\Delta u$  in equation (5.9) of theorem 5.2.2. For an arbitrary function  $v$  of  $H_0^1(\Omega)$  we have not given (even a weak) meaning to its Laplacian  $\Delta v$ . Conversely, for the solution  $u \in H_0^1(\Omega)$  of the variational formulation (5.5), we have shown that  $\Delta u$  belongs to  $L^2(\Omega)$ . •

For the boundary value problem (5.1) to be well-posed (in the sense of Hadamard; see definition 1.5.3), we must, in addition to existence and uniqueness of the solution, show that the solution depends continuously on the data. This is an immediate consequence of the Lax–Milgram theorem 3.3.1 but we shall present a new statement and a new proof.

**Proposition 5.2.5** *Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$ , and let  $f \in L^2(\Omega)$ . The mapping which takes  $f \in L^2(\Omega)$  to the unique solution  $u \in H_0^1(\Omega)$  of the variational formulation of (5.1) is linear and continuous from  $L^2(\Omega)$  into  $H^1(\Omega)$ . In particular, there exists a constant  $C > 0$  such that, for all  $f \in L^2(\Omega)$ , we have*

$$\|u\|_{H^1(\Omega)} \leq C\|f\|_{L^2(\Omega)}. \quad (5.10)$$

**Remark 5.2.6** The inequality (5.10) is what we call an **energy estimate**. It guarantees that the energy of the solution is controlled by that of the data. Energy estimates are very natural from a physical viewpoint and very useful from a mathematical viewpoint. •

**Proof.** The linearity of  $f \rightarrow u$  is obvious. To obtain the continuity we take  $v = u$  in the variational formulation (5.5)

$$\int_{\Omega} |\nabla u|^2 dx = \int_{\Omega} f u dx.$$

We obtain an upper bound on the term on the right with the help of the Cauchy–Schwarz inequality, and a lower bound on that on the left by the coercivity of the bilinear form

$$\nu \|u\|_{H^1(\Omega)}^2 \leq \|f\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \|u\|_{H^1(\Omega)},$$

from which we deduce the result.  $\square$

We have already said that the variational formulation often has a physical interpretation (this is, for example, the principle of virtual work in mechanics). In fact, the solution of the variational formulation (5.5) attains the minimum of an energy (very natural in physics or mechanics). The following result is an immediate application of proposition 3.3.4.

**Proposition 5.2.7** *Let  $J(v)$  be the energy defined for  $v \in H_0^1(\Omega)$  by*

$$J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx. \quad (5.11)$$

*Let  $u \in H_0^1(\Omega)$  be the unique solution of the variational formulation (5.5). Then  $u$  is also the unique minimum of the energy, that is,*

$$J(u) = \min_{v \in H_0^1(\Omega)} J(v).$$

*Conversely, if  $u \in H_0^1(\Omega)$  is a minimum of the energy  $J(v)$ , then  $u$  is the unique solution of the variational formulation (5.5).*

**Remark 5.2.8** Proposition 5.2.7 relies crucially on the fact that the bilinear form of the variational formulation is symmetric. If this is not the case, the solution of the variational formulation does not minimize the energy (see the counterexample of exercise 5.2.3).

Often the physical origin of Laplacian is in fact the search for minima of the energy  $J(v)$ . It is remarkable that this minimization problem needs the solution  $u$  to have less regularity than the partial differential equation (only one derivative allows us to define  $J(u)$  while we need two for  $\Delta u$ ). This observation confirms the ‘natural’ character of the variational formulation to analyse a PDE.  $\bullet$

**Exercise 5.2.1** With the help of the variational approach show the existence and uniqueness of the solution of

$$\begin{cases} -\Delta u + u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (5.12)$$

where  $\Omega$  is an arbitrary open set of the space  $\mathbb{R}^N$ , and  $f \in L^2(\Omega)$ . Show in particular that the addition of a term of order zero to the Laplacian allows us to ignore the hypothesis that  $\Omega$  is bounded.

**Exercise 5.2.2** Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$ . With the help of the variational approach show the existence and uniqueness of the solution of the following convection–diffusion problem

$$\begin{cases} V \cdot \nabla u - \Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (5.13)$$

where  $f \in L^2(\Omega)$  and  $V$  is a regular vector valued function such that  $\operatorname{div} V = 0$  in  $\Omega$ .

**Exercise 5.2.3** We take the notation and hypotheses of exercise 5.2.2. Show that every  $v \in H_0^1(\Omega)$  satisfies

$$\int_{\Omega} v V \cdot \nabla v \, dx = 0.$$

Show that the solution of the variational formulation of the convection–diffusion problem does not minimize the energy

$$J(v) = \frac{1}{2} \int_{\Omega} (|\nabla v|^2 + v V \cdot \nabla v) \, dx - \int_{\Omega} f v \, dx,$$

in  $H_0^1(\Omega)$ .

The ‘uniqueness’ part of theorem 5.2.2 is useful to show some symmetry properties as the following exercise indicates.

**Exercise 5.2.4** We consider again the boundary value problem (5.1). We assume that the open set  $\Omega$  is symmetric with respect to the hyperplane  $x_N = 0$ ; likewise for the data  $f$  (that is,  $f(x', x_N) = f(x', -x_N)$ ). Show that the solution of (5.1) has the same symmetry. Show that (5.1) is equivalent to a boundary value problem posed over  $\Omega^+ = \Omega \cap \{x_N > 0\}$  with a Neumann boundary condition on  $\Omega \cap \{x_N = 0\}$ .

**Remark 5.2.9** Until now we have assumed that the right-hand side  $f$  of (5.1) belongs to  $L^2(\Omega)$ , but many of the results remain true if we only assume that  $f \in H^{-1}(\Omega)$  (a less regular space of ‘functions’). The two first stages remain identical if we replace the usual integral  $\int_{\Omega} f v \, dx$  by the duality pairing  $\langle f, v \rangle_{H^{-1}, H_0^1(\Omega)}$  (in particular,  $L(v)$  is still a continuous linear form over  $H_0^1(\Omega)$ ). In the third stage, the equation  $-\Delta u = f$  no longer holds in the sense of the equality between the elements of  $H^{-1}(\Omega)$  (nor almost everywhere in  $\Omega$ ).

This mathematical refinement can correspond to a pertinent physical model. Take as an example the case of a right-hand side concentrated on a hypersurface rather than distributed over all of  $\Omega$ . Let  $\Gamma$  be a regular hypersurface (a manifold of dimension  $N-1$ ) included in  $\Omega$ . To model a concentrated source term over  $\Gamma$ , we take  $\tilde{f} \in L^2(\Gamma)$  and we define  $f \in H^{-1}(\Omega)$  by

$$\langle f, v \rangle_{H^{-1}, H_0^1(\Omega)} = \int_{\Gamma} \tilde{f} v \, ds,$$

which is a continuous linear form over  $H_0^1(\Omega)$  thanks to the trace theorem 4.3.13. •

**Remark 5.2.10** In (5.1) we have considered ‘homogeneous’, that is, zero, Dirichlet boundary conditions but we can also treat the case of nonhomogeneous boundary conditions. Consider the boundary value problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = u_0 & \text{on } \partial\Omega, \end{cases} \quad (5.14)$$

where  $u_0$  is the trace over  $\partial\Omega$  of a function of  $H^1(\Omega)$ , (again denoted by  $u_0$ ). To analyse (5.14) we set  $u = u_0 + \tilde{u}$ , and we look for the solution of

$$\begin{cases} -\Delta \tilde{u} = \tilde{f} = f + \Delta u_0 & \text{in } \Omega \\ \tilde{u} = 0 & \text{on } \partial\Omega. \end{cases} \quad (5.15)$$

Following remark 5.2.9 we can solve (5.15) by the variational approach as  $\tilde{f}$  belongs to  $H^{-1}(\Omega)$ . In effect,

$$\langle \tilde{f}, v \rangle_{H^{-1}, H_0^1(\Omega)} = \int_{\Omega} f v \, dx - \int_{\Omega} \nabla u_0 \cdot \nabla v \, dx$$

is a continuous linear form over  $H_0^1(\Omega)$ . •

## 5.2.2 Neumann boundary conditions

We consider the following boundary value problem

$$\begin{cases} -\Delta u + u = f & \text{in } \Omega \\ \frac{\partial u}{\partial n} = g & \text{on } \partial\Omega \end{cases} \quad (5.16)$$

where  $\Omega$  is an open set (not necessarily bounded) of the space  $\mathbb{R}^N$ ,  $f \in L^2(\Omega)$  and  $g \in L^2(\partial\Omega)$ . The equation (5.16) is a variant of the Laplacian where we have added a term of order zero to avoid (in the first instance) a difficulty that we will deal with later in theorem 5.2.18. The variational approach to studying (5.16) is appreciably different from that presented in the preceding section in the treatment of the boundary conditions. This is why we again detail the three stages of the approach.

### Stage 1: Establishment of a variational formulation

To find the variational formulation we multiply equation (5.16) by a regular test function  $v$  and we integrate by parts assuming that the solution  $u$  is sufficiently regular so that all the calculations are valid. Green’s formula (4.22) (see also (3.7)) gives

$$\begin{aligned} \int_{\Omega} f(x)v(x) \, dx &= \int_{\Omega} (-\Delta u(x) + u(x))v(x) \, dx \\ &= \int_{\Omega} (\nabla u(x) \cdot \nabla v(x) + u(x)v(x)) \, dx - \int_{\partial\Omega} \frac{\partial u}{\partial n}(x)v(x) \, ds \\ &= \int_{\Omega} (\nabla u(x) \cdot \nabla v(x) + u(x)v(x)) \, dx - \int_{\partial\Omega} g(x)v(x) \, ds. \end{aligned} \quad (5.17)$$

We have used the Neumann boundary conditions in (5.17) and it is not necessary to include this in the choice of Hilbert space  $V$ . For the first term and the two last terms of (5.17) to have a meaning it is sufficient to take  $V = H^1(\Omega)$  (we use the trace theorem 4.3.13 to justify the boundary integral).

In conclusion, the variational formulation proposed for (5.16) is: find  $u \in H^1(\Omega)$  such that

$$\int_{\Omega} (\nabla u \cdot \nabla v + uv) dx = \int_{\partial\Omega} gv ds + \int_{\Omega} fv dx \quad \forall v \in H^1(\Omega). \quad (5.18)$$

The following stages justify the choice of (5.18).

**Remark 5.2.11** The principal difference between the variational formulation (5.18) for Neumann boundary conditions and (5.5) for Dirichlet boundary conditions is that the Dirichlet condition is included in the choice of the space while the Neumann condition appears in the linear form but not in the space. The Dirichlet condition is called **essential** (or explicit) since it is forced by the space, while the Neumann condition is called **natural** (or implicit) since it comes from the integration by parts which leads to the variational formulation. •

## Stage 2: Solution of the variational formulation

In this second stage we verify that the variational formulation (5.18) has a unique solution. For this we use the Lax–Milgram theorem 3.3.1 whose hypotheses we verify with the notation

$$a(u, v) = \int_{\Omega} (\nabla u \cdot \nabla v + uv) dx \quad \text{and} \quad L(v) = \int_{\partial\Omega} gv ds + \int_{\Omega} fv dx.$$

By using the Cauchy–Schwarz inequality and with the help of the trace theorem 4.3.13, we clearly see that  $a$  is a continuous bilinear form over  $H^1(\Omega)$  and that  $L$  is a continuous linear form over  $H^1(\Omega)$ . In addition, the bilinear form  $a$  is obviously coercive (this is why we have added a term of order zero to the Laplacian) since

$$a(v, v) = \|v\|_{H^1(\Omega)}^2 \quad \forall v \in H^1(\Omega).$$

As  $H^1(\Omega)$  is a Hilbert space (see proposition 4.3.2), all the hypotheses of the Lax–Milgram theorem 3.3.1 are satisfied and we can therefore conclude that there exists a unique solution  $u \in H^1(\Omega)$  of the variational formulation (5.18).

**Remark 5.2.12** To analyse the boundary value problem (5.16) in the case where  $g = 0$ , we might be tempted to include the Neumann boundary condition in the Hilbert space  $V$ . It is not possible to choose  $V = \{v \in H^1(\Omega), \frac{\partial v}{\partial n} = 0 \text{ on } \partial\Omega\}$  since, for a function  $v \in H^1(\Omega)$ ,  $\frac{\partial v}{\partial n}$  does not have a meaning over  $\partial\Omega$ . In effect,  $\nabla v$  is only a function of  $L^2(\Omega)$  (component by component) and we know there is no idea of a trace  $\partial\Omega$  for the functions of  $L^2(\Omega)$ . We could choose  $V = \{v \in H^2(\Omega), \frac{\partial v}{\partial n} = 0 \text{ on } \partial\Omega\}$

which is a closed subspace of  $H^2(\Omega)$  from the trace theorem 4.3.28. But with this last choice a new difficulty emerges: the bilinear form  $a$  is not coercive over  $V$  and we cannot apply the Lax–Milgram theorem. There is therefore no way of taking into account the Neumann boundary conditions in the choice of the Hilbert space. •

**Stage 3:** Equivalence with the equation.

We now interpret the variational formulation (5.18) to verify that we have solved the boundary value problem (5.16), in a sense to be made precise. We shall assume that the data are regular (see remark 5.2.15 if this is not the case). More precisely we shall assume that we are in the position to apply the following regularity lemma (see Section 5.2.4 for similar results).

**Lemma 5.2.13** *Let  $\Omega$  be an open set regular of class  $\mathcal{C}^1$  of  $\mathbb{R}^N$ . Take  $f \in L^2(\Omega)$  and let  $g$  be the trace over  $\partial\Omega$  of a function of  $H^1(\Omega)$ . Then the solution  $u$  of the variational formulation (5.18) belongs to  $H^2(\Omega)$ .*

Thanks to lemma 5.2.13 we can use the Green's formula of theorem 4.3.30

$$\int_{\Omega} \Delta u(x) v(x) dx = - \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \frac{\partial u}{\partial n}(x) v(x) ds. \quad (5.19)$$

which is valid for  $u \in H^2(\Omega)$  and  $v \in H^1(\Omega)$ . Recall that the boundary integral in (5.19) has a meaning because of the trace theorem 4.3.28 which confirms that for  $u \in H^2(\Omega)$  the normal derivative  $\frac{\partial u}{\partial n}$  has a meaning in  $L^2(\partial\Omega)$ . We deduce then from (5.18) and (5.19) that, for all  $v \in H^1(\Omega)$ ,

$$\int_{\Omega} (\Delta u - u + f) v dx = \int_{\partial\Omega} \left( g - \frac{\partial u}{\partial n} \right) v ds. \quad (5.20)$$

If we take  $v \in C_c^\infty(\Omega) \subset H^1(\Omega)$  in (5.20), the boundary term disappears and we deduce, from corollary 4.2.2, that  $\Delta u - u + f = 0$  in  $L^2(\Omega)$ , therefore almost everywhere in  $\Omega$ . Consequently, the left-hand side of (5.20) is zero, therefore

$$\int_{\partial\Omega} \left( g - \frac{\partial u}{\partial n} \right) v ds = 0 \quad \forall v \in H^1(\Omega).$$

Now the image of  $H^1(\Omega)$  by the trace mapping is dense in  $L^2(\partial\Omega)$  (see remark 4.3.17), which implies that  $g - \frac{\partial u}{\partial n} = 0$  in  $L^2(\partial\Omega)$ , and therefore almost everywhere on  $\partial\Omega$ . In conclusion, we have proved the following result.

**Theorem 5.2.14** *Let  $\Omega$  be an open set regular of class  $\mathcal{C}^1$  of  $\mathbb{R}^N$ . Take  $f \in L^2(\Omega)$  and let  $g$  be the trace over  $\partial\Omega$  of a function of  $H^1(\Omega)$ . There exists a unique solution  $u \in H^1(\Omega)$  of the variational formulation (5.18). Further,  $u$  belongs to  $H^2(\Omega)$  and is the solution of (5.16) in the sense of*

$$-\Delta u + u = f \text{ almost everywhere in } \Omega, \quad \frac{\partial u}{\partial n} = g \text{ almost everywhere over } \partial\Omega.$$

**Exercise 5.2.5** Show that the unique solution  $u \in H^1(\Omega)$  of the variational formulation (5.18) satisfies the following energy estimate

$$\|u\|_{H^1(\Omega)} \leq C (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}),$$

where  $C > 0$  is a constant which does not depend on  $u, f$  and  $g$ .

As in the preceding section, in practice (and when the context does not lead to confusion) we say that the unique solution  $u \in H^1(\Omega)$  of the variational formulation (5.18) is the unique **weak solution** of the boundary value problem (5.16).

**Remark 5.2.15 (delicate)** When the open set  $\Omega$  is not regular and  $g$  is only a function of  $L^2(\partial\Omega)$ , we have also proved the existence of a unique solution  $u \in H^1(\Omega)$  of the variational formulation (5.18). To show that this solution satisfies the equation  $-\Delta u + u = f$  almost everywhere in  $\Omega$ , we must use an argument which is a little more complicated. Denoting by  $\sigma = \nabla u \in L^2(\Omega)^N$ , we deduce from (5.18) that, for all  $v \in C_c^\infty(\Omega)$ ,

$$\left| \int_{\Omega} \sigma \cdot \nabla v \, dx \right| \leq \left| \int_{\Omega} uv \, dx \right| + \left| \int_{\Omega} fv \, dx \right| \leq C \|v\|_{L^2(\Omega)}$$

which is none other than the criterion for the existence of a weak divergence of  $\sigma$  in  $L^2(\Omega)$  (see definition 4.2.6 and the lemma 4.2.7). We therefore have  $\operatorname{div} \sigma \in L^2(\Omega)$  and

$$\int_{\Omega} (\operatorname{div} \sigma - u + f) v \, dx = 0 \quad \forall v \in C_c^\infty(\Omega),$$

which implies, from corollary 4.2.2, that  $-\operatorname{div} \sigma = -\Delta u = f - u$  in  $L^2(\Omega)$ .

We can also recover the Neumann boundary conditions, in a very weak sense, if the open set (but not  $g$ ) is regular. To do this we use the space  $H(\operatorname{div})$  introduced in section 4.4.2 and defined by  $H(\operatorname{div}) = \{\sigma \in L^2(\Omega)^N \text{ such that } \operatorname{div} \sigma \in L^2(\Omega)\}$ . Theorem 4.4.7 confirms that, if  $\Omega$  is an open set regular of class  $C^1$ , we have the following integration by parts formula

$$\int_{\Omega} \operatorname{div} \sigma v \, dx + \int_{\Omega} \sigma \cdot \nabla v \, dx = \langle \sigma \cdot n, v \rangle_{H^{-1/2}, H^{1/2}(\partial\Omega)}, \quad (5.21)$$

for  $v \in H^1(\Omega)$  and  $\sigma \in H(\operatorname{div})$ . The term on the right of (5.21) denotes the duality pairing between  $H^{1/2}(\partial\Omega)$  and its dual, denoted  $H^{-1/2}(\partial\Omega)$ . If  $\sigma$  and  $v$  are regular functions, the 'bad' term is only the usual boundary integral  $\int_{\partial\Omega} v \sigma \cdot n \, ds$ . If not, formula (5.21) gives a meaning to  $\sigma \cdot n$  over  $\partial\Omega$  (as an element of the dual  $H^{-1/2}(\partial\Omega)$ ) for  $\sigma \in H(\operatorname{div})$ .

If we apply this result to the solution of (5.18) (with  $\sigma = \nabla u$ ), we deduce that the Neumann boundary condition is satisfied as an equation between elements of the dual  $H^{-1/2}(\partial\Omega)$  (as  $g \in L^2(\partial\Omega) \subset H^{-1/2}(\partial\Omega)$ ). This argument is reasonably complicated, and in practice we shall be content with saying that the variational formulation contains a generalization of the Neumann boundary conditions, and in practice we shall continue to write **formally** that  $\frac{\partial u}{\partial n} = g$  on  $\partial\Omega$ . •

As in the preceding subsection, we can show that the solution of (5.16) minimizes an energy. We remark that, if  $g = 0$ , then the energy (5.22) is the same as that of



(5.11), defined for the Laplacian with Dirichlet boundary conditions. Nevertheless, their minima are in general not the same as we minimize over two different spaces, that is,  $H_0^1(\Omega)$  and  $H^1(\Omega)$ . The following proposition is an immediate application of proposition 3.3.4.

**Proposition 5.2.16** *Let  $J(v)$  be the energy defined for  $v \in H^1(\Omega)$  by*

$$J(v) = \frac{1}{2} \int_{\Omega} (|\nabla v|^2 + |v|^2) dx - \int_{\Omega} f v dx - \int_{\partial\Omega} g v ds. \quad (5.22)$$

*Let  $u \in H^1(\Omega)$  be the unique solution of the variational formulation (5.18). Then  $u$  is also the unique minimum of the energy, that is,*

$$J(u) = \min_{v \in H^1(\Omega)} J(v).$$

*Conversely, if  $u \in H^1(\Omega)$  is a minimum of the energy  $J(v)$ , then  $u$  is the unique solution of the variational formulation (5.18).*

**Exercise 5.2.6** We assume that  $\Omega$  is a regular open bounded set of class  $\mathcal{C}^1$ . With the help of the variational approach, show the existence and uniqueness of the solution of the Laplacian with a Fourier boundary condition

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ \frac{\partial u}{\partial n} + u = g & \text{on } \partial\Omega \end{cases} \quad (5.23)$$

where  $f \in L^2(\Omega)$  and  $g$  is the trace over  $\partial\Omega$  of a function of  $H^1(\Omega)$ . We shall show the following inequality (which generalizes the Poincaré inequality)

$$\|v\|_{L^2(\Omega)} \leq C (\|v\|_{L^2(\partial\Omega)} + \|\nabla v\|_{L^2(\Omega)}) \quad \forall v \in H^1(\Omega).$$

**Exercise 5.2.7** We assume that  $\Omega$  is an open bounded connected set. With the help of the variational approach show the existence and uniqueness of the solution of the Laplacian with mixed boundary conditions

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega_N \\ u = 0 & \text{on } \partial\Omega_D \end{cases} \quad (5.24)$$

where  $f \in L^2(\Omega)$ , and  $(\partial\Omega_N, \partial\Omega_D)$  is a partition of  $\partial\Omega$  such that the surface measures of  $\partial\Omega_N$  and  $\partial\Omega_D$  are nonzero (see Figure 4.1). (Use remark 4.3.18.)

We return now to a true Laplacian operator (without the addition of a term of order zero as in (5.16)) and we consider the boundary value problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ \frac{\partial u}{\partial n} = g & \text{on } \partial\Omega \end{cases} \quad (5.25)$$

where  $\Omega$  is an open bounded connected set of the space  $\mathbb{R}^N$ ,  $f \in L^2(\Omega)$  and  $g \in L^2(\partial\Omega)$ . The new difficulty in (5.25) with respect to (5.16) is that there only exists a solution if the data  $f$  and  $g$  satisfy a **compatibility condition**. In effect, it is easy to see that if there exists a solution  $u \in H^2(\Omega)$ , then integrating the equation over  $\Omega$  (or using Green's formula (4.22)) we must have

$$\int_{\Omega} f(x) dx + \int_{\partial\Omega} g(x) ds = 0. \quad (5.26)$$

We remark also that if  $u$  is a solution then  $u + C$ , with  $C \in \mathbb{R}$ , is also a solution. In fact, (5.26) is a necessary and sufficient condition for the existence of a solution in  $H^1(\Omega)$ , unique up to the addition of an arbitrary constant. We remark that, if the open set  $\Omega$  is not connected, then we must write (5.26) for each connected component of  $\Omega$  and the uniqueness of the solution holds up to the addition of an arbitrary constant in each connected component (with these modifications all the results which follow remain valid).

**Remark 5.2.17** Physically, the compatibility condition (5.26) is interpreted as an **equilibrium condition**:  $f$  corresponds to a volume source and  $g$  a flux entering at the boundary. So that there exists a stationary or equilibrium state (that is, a solution of (5.25)), these two terms must balance exactly. Likewise, the uniqueness 'up to a constant' corresponds to the absence of a reference scale on which to measure the values of  $u$  (like for temperature, for example). •

**Theorem 5.2.18** *Let  $\Omega$  be a regular open bounded connected set of class  $\mathcal{C}^1$  of  $\mathbb{R}^N$ . Take  $f \in L^2(\Omega)$  and  $g \in L^2(\partial\Omega)$  which satisfy the compatibility condition (5.26). There exists a weak solution  $u \in H^1(\Omega)$  of (5.25), unique up to an additive constant.*

**Proof.** To find the variational formulation we proceed as for equation (5.16). A similar calculation leads to

$$\int_{\Omega} \nabla u \cdot \nabla v dx = \int_{\partial\Omega} gv ds + \int_{\Omega} fv dx$$

for every regular test function  $v$ . To give a meaning to all the terms of this equation, we could choose  $H^1(\Omega)$  as the Hilbert space  $V$ , but we could not show the coercivity of the bilinear form. This difficulty is intimately linked with the fact that, if  $u$  is a solution, then  $u + C$  is also a solution. To avoid this disadvantage, we work only with functions of average zero. In other words, we set

$$V = \left\{ v \in H^1(\Omega), \int_{\Omega} v(x) dx = 0 \right\}$$

and the variational formulation of (5.25) is:

$$\text{find } u \in V \text{ such that } \int_{\Omega} \nabla u \cdot \nabla v dx = \int_{\partial\Omega} gv ds + \int_{\Omega} fv dx \quad \forall v \in V. \quad (5.27)$$

We can equally choose the quotient space  $V = H^1(\Omega)/\mathbb{R}$  (the elements of  $H^1(\Omega)/\mathbb{R}$  are the classes of functions of  $H^1(\Omega)$  equal up to a constant).

To be able to apply the Lax–Milgram theorem to the variational formulation (5.27), the only delicate hypothesis to verify is the coercivity of the bilinear form. This is obtained thanks to a generalization of the Poincaré inequality, known as the Poincaré–Wirtinger inequality: if  $\Omega$  is bounded and connected, there exists a constant  $C > 0$  such that, for all  $v \in H^1(\Omega)$ ,

$$\|v - m(v)\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)} \quad \text{with } m(v) = \frac{\int_{\Omega} v \, dx}{\int_{\Omega} 1 \, dx}. \quad (5.28)$$

The inequality (5.28) is proved by contradiction like the second proof of proposition 4.3.10 (we leave this as an exercise to the reader). As  $m(v) = 0$  for all  $v \in V$ , (5.28) implies that  $\|\nabla v\|_{L^2(\Omega)}$  is a norm in  $V$ , equivalent to the usual norm  $\|v\|_{H^1(\Omega)}$ , and therefore that the bilinear form is coercive over  $V$ .

Finally, to show that the unique solution of (5.27) is a solution of the boundary value problem (5.25), we proceed as we did in the proof of theorem 5.2.14. We thus obtain for all  $v \in V$ ,

$$\int_{\Omega} (\Delta u + f)v \, dx = \int_{\partial\Omega} \left( g - \frac{\partial u}{\partial n} \right) v \, ds. \quad (5.29)$$

However, for  $w \in H^1(\Omega)$ , the function  $v = w - m(w)$  belongs to  $V$ . By choosing such a function in (5.29), and rearranging the terms and using the compatibility condition (5.26) as well as the equation  $\int_{\Omega} \Delta u \, dx = \int_{\partial\Omega} \frac{\partial u}{\partial n} \, ds$ , we deduce from (5.29)

$$\int_{\Omega} (\Delta u + f)w \, dx = \int_{\partial\Omega} \left( g - \frac{\partial u}{\partial n} \right) w \, ds \quad \forall w \in H^1(\Omega).$$

We can therefore conclude as usual that  $u$  satisfies the boundary value problem (5.25).  $\square$

**Exercise 5.2.8** Show the Poincaré–Wirtinger inequality (5.28).

**Exercise 5.2.9** We assume that  $\Omega$  is a regular open bounded connected set. Take  $f \in L^2(\Omega)$ . We consider the following variational formulation: find  $u \in H^1(\Omega)$  such that

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx + \left( \int_{\Omega} u \, dx \right) \left( \int_{\Omega} v \, dx \right) = \int_{\Omega} f v \, dx \quad \forall v \in H^1(\Omega).$$

Show the existence and uniqueness of the solution of this variational formulation. Which boundary value problem have we solved? In particular, if we assume that  $\int_{\Omega} f \, dx = 0$ , which problem already studied have we recovered?

### 5.2.3 Variable coefficients

In the two preceding sections we have considered boundary value problems for the Laplacian operator. We can easily generalize the results obtained to more general operators, so-called second order elliptic with **variable coefficients**. This type of problem arises from the modelling of heterogeneous media. If we return to the example of the conduction of heat (detailed in Chapter 1), in a heterogeneous medium the conductivity  $k(x)$  is a function which varies across the domain. In this case, we consider the boundary value problem

$$\begin{cases} -\operatorname{div}(k\nabla u) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (5.30)$$

where  $\Omega$  is an open bounded set of the space  $\mathbb{R}^N$ , and  $f \in L^2(\Omega)$ . Of course, if  $k(x) \equiv 1$ , we recover the Laplacian. It is easy to generalize theorem 5.2.2.

**Proposition 5.2.19** *Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$ . Let  $f \in L^2(\Omega)$ . We assume that the coefficient  $k(x)$  is a measurable function and that there exist two strictly positive constants  $0 < k^- \leq k^+$  such that*

$$0 < k^- \leq k(x) \leq k^+ \text{ almost everywhere } x \in \Omega. \quad (5.31)$$

*Then, there exists a unique (weak) solution  $u \in H_0^1(\Omega)$  of (5.30).*

**Proof.** To find the variational formulation we multiply the equation (5.39) by a test function  $v$  and we integrate by parts using Green's formula of exercise 3.2.1

$$\int_{\Omega} \operatorname{div} \sigma(x) v(x) dx = - \int_{\Omega} \sigma(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \sigma(x) \cdot n(x) v(x) ds,$$

with  $\sigma = k\nabla u$ . To take account of the Dirichlet boundary conditions we choose  $H_0^1(\Omega)$  as the Hilbert space, and we find the variational formulation of (5.30):

$$\text{find } u \in H_0^1(\Omega) \text{ such that } \int_{\Omega} k\nabla u \cdot \nabla v dx = \int_{\Omega} f v dx \quad \forall v \in H_0^1(\Omega). \quad (5.32)$$

Thanks to hypothesis (5.31) we know that the bilinear form of (5.32) is continuous

$$\left| \int_{\Omega} k\nabla u \cdot \nabla v dx \right| \leq k^+ \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)},$$

and that it is coercive

$$\int_{\Omega} k\nabla u \cdot \nabla u dx \geq k^- \int_{\Omega} |\nabla u|^2 dx \geq \nu \|u\|_{H_0^1(\Omega)},$$

with  $\nu > 0$  thanks to the Poincaré inequality. We can therefore apply the Lax–Milgram theorem, which proves the existence and uniqueness of the solution of the variational formulation (5.32). To show that this variational solution is also a solution of the boundary value problem (5.30), we proceed as in the proof of theorem 5.2.2.  $\square$

**Remark 5.2.20** It is very important to write the equation (5.30) in divergence form: we could *a priori* also write it

$$-\operatorname{div}(k\nabla u) = -k\Delta u - \nabla k \cdot \nabla u,$$

but this last form only has a meaning if the coefficient  $k$  is differentiable. Conversely, the equation (5.30) has a meaning even if  $k$  is discontinuous. •

In fact, when the equation (5.30) is in divergence form, its interpretation ‘in the weak sense’ contains more information than its classical statement. More precisely, the idea of weak divergence implicitly contains what we call the **transmission boundary conditions** between two subdomains occupied by two materials of different conductivity. We consider an example where  $(\Omega_1, \Omega_2)$  is a partition of  $\Omega$  over which  $k(x)$  is piecewise constant

$$k(x) = k_i > 0 \quad \text{for } x \in \Omega_i, i = 1, 2. \quad (5.33)$$

We denote by  $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$  the interface (assumed regular and included in  $\Omega$ ) between  $\Omega_1$  and  $\Omega_2$  (see Figure 5.1), and  $u_i = u|_{\Omega_i}$  the restriction of the solution  $u$  to  $\Omega_i$ .

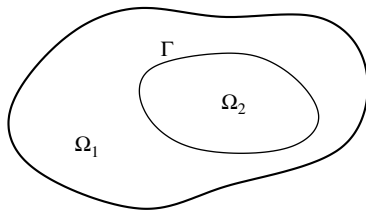


Figure 5.1. Interface between two subdomains and transmission condition.

**Lemma 5.2.21** *Under the hypothesis (5.33) the problem (5.30) is equivalent to*

$$\begin{cases} -k_i \Delta u_i = f & \text{in } \Omega_i, i = 1, 2, \\ u_1 = 0 & \text{on } \partial\Omega, \\ u_1 = u_2 & \text{on } \Gamma, \\ k_1 \nabla u_1 \cdot n = k_2 \nabla u_2 \cdot n & \text{on } \Gamma. \end{cases} \quad (5.34)$$

The two last lines of (5.34) are called **transmission boundary conditions** on the interface  $\Gamma$ .

**Proof.** If  $u \in H_0^1(\Omega)$  is a solution of (5.30), by application of the trace theorem 4.3.13, we must have  $u_1 = u_2$  on  $\Gamma$ . If we set  $\sigma = k\nabla u$  and  $\sigma_i = \sigma|_{\Omega_i} = k_i \nabla u_i$  its restriction to  $\Omega_i$ , we know that  $\sigma$ , as well as its divergence, belong to  $L^2(\Omega)$ . Then,

from theorem 4.4.7, the normal component  $\sigma \cdot n$  has a meaning on  $\Gamma$  and we must have  $\sigma_1 \cdot n = \sigma_2 \cdot n$  on  $\Gamma$ .

Conversely, we construct a variational formulation of (5.34) to show that it has a unique solution which coincides with the solution  $u$  of (5.30). We look for  $u_i \in H^1(\Omega_i)$  and we multiply each equation in  $\Omega_i$  by the same test function  $v \in H_0^1(\Omega)$ . By integrating by parts and summing we obtain

$$\int_{\Omega_1} \nabla u_1 \cdot \nabla v \, dx + \int_{\Omega_2} \nabla u_2 \cdot \nabla v \, dx + \int_{\Gamma} \left( k_1 \frac{\partial u_1}{\partial n_1} + k_2 \frac{\partial u_2}{\partial n_2} \right) v \, ds = \int_{\Omega_1} f v \, dx + \int_{\Omega_2} f v \, dx. \quad (5.35)$$

Since  $n_1 = -n_2$  the integral on the interface  $\Gamma$  disappears because of the transmission boundary conditions. On the other hand, if  $u$  is defined as  $u_1$  in  $\Omega_1$  and  $u_2$  in  $\Omega_2$ , the transmission condition  $u_1 = u_2$  on  $\Gamma$  implies that  $u \in H^1(\Omega)$  from lemma 4.3.19. Consequently, (5.35) is nothing other than the variational formulation (5.32).  $\square$

**Exercise 5.2.10** Let  $\Omega$  be an open bounded set and  $K$  a compact connected set of  $\mathbb{R}^N$  included in  $\Omega$  (we assume that  $\Omega \setminus K$  is regular). Take  $f \in L^2(\Omega)$ . We consider a conduction problem in  $\Omega$  where  $K$  is a perfect conductor, that is, the unknown  $u$  (the temperature or the electrical potential, for example) is constant in  $K$  (this constant is also unknown). We assume that there is no source term in  $K$ . This problem is modelled by

$$\begin{cases} -\Delta u = f & \text{in } \Omega \setminus K \\ u = C & \text{on } \partial K \\ \int_{\partial K} \frac{\partial u}{\partial n} \, ds = 0 & \text{on } \partial K \\ u = 0 & \text{on } \partial \Omega, \end{cases}$$

where  $C$  is an unknown constant to be determined. Find a variational formulation of this boundary value problem and show the existence and uniqueness of a solution  $(u, C)$ .

We can again generalize the above to more general operators with tensorial coefficients  $A(x) = (a_{ij}(x))_{1 \leq i, j \leq N}$ . We assume that the matrix  $A$  is uniformly positive definite over  $\Omega$  (or coercive, or elliptic), that is, there exists a constant  $\alpha > 0$  such that, almost everywhere in  $\Omega$ ,

$$A(x)\xi \cdot \xi = \sum_{i,j=1}^N a_{ij}(x)\xi_i\xi_j \geq \alpha|\xi|^2 \quad \text{for all } \xi \in \mathbb{R}^N, \quad (5.36)$$

and that it is uniformly bounded, that is, there exists a constant  $\beta > 0$  such that, almost everywhere in  $\Omega$ ,

$$|A(x)\xi| \leq \beta|\xi| \quad \text{for all } \xi \in \mathbb{R}^N. \quad (5.37)$$

We then define the operator

$$-\operatorname{div}(A\nabla \cdot) = - \sum_{i,j=1}^N \frac{\partial}{\partial x_i} \left( a_{ij}(x) \frac{\partial}{\partial x_j} \cdot \right), \quad (5.38)$$

and we consider the boundary value problem

$$\begin{cases} -\operatorname{div}(A\nabla u) = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (5.39)$$

Several physical motivations lead to models of the type of (5.39). In the example of heat conduction, an anisotropic medium (where the conductivity is not the same in all directions) is characterized by a symmetric conductivity matrix  $A(x)$  (not proportional to the identity). The case of matrices  $A(x)$  which are nonsymmetric corresponds, for example, to taking account of a convection effect. In effect, if we decompose  $A = A^s + A^a$  into its symmetric part  $A^s = (A + A^t)/2$  and its antisymmetric part  $A^a = (A - A^t)/2$ , a simple calculation shows that

$$-\operatorname{div}(A\nabla u) = -\operatorname{div}(A^s\nabla u) + V \cdot \nabla u \quad \text{with } V_j(x) = \sum_{i=1}^N \frac{1}{2} \frac{\partial(a_{ji} - a_{ij})}{\partial x_i}(x),$$

where  $V$  is interpreted as a convection velocity.

**Exercise 5.2.11** Show under the hypotheses (5.36) and (5.37) that (5.39) has a unique (weak) solution  $u \in H_0^1(\Omega)$  if  $f \in L^2(\Omega)$ .

We can replace the Dirichlet boundary condition in (5.39) by a Neumann boundary condition which, for the operator (5.38), is written

$$\frac{\partial u}{\partial n_A} = \left( A(x)\nabla u \right) \cdot n = \sum_{i,j=1}^N a_{ij}(x) \frac{\partial u}{\partial x_j} n_i = 0 \quad \text{on } \partial\Omega,$$

where  $\frac{\partial u}{\partial n_A}$  is the so-called conormal derivative of  $u$  associated with the operator (5.38). This Neumann condition is very natural from the point of view of physics since, if we introduce the flux  $\sigma = A\nabla u$ , it implies that the normal component of the flux is zero on the boundary  $\sigma \cdot n = 0$  over  $\partial\Omega$ .

**Exercise 5.2.12** For  $f \in L^2(\Omega)$ ,  $g \in L^2(\partial\Omega)$ , show the existence and uniqueness of the solution of

$$\begin{cases} -\operatorname{div}(A\nabla u) + u = f & \text{in } \Omega, \\ \frac{\partial u}{\partial n_A} = g & \text{on } \partial\Omega. \end{cases}$$

## 5.2.4 Qualitative properties

In this section we study some qualitative properties of solutions of the Laplacian with Dirichlet boundary conditions. In all this section,  $\Omega$  is an open bounded set of  $\mathbb{R}^N$  and  $f \in L^2(\Omega)$ . Theorem 5.2.2 gives a unique solution  $u \in H_0^1(\Omega)$  of

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (5.40)$$

### Maximum principle

We start by recovering this property discovered in Chapter 1 (thanks to some explicit formulas in  $N = 1$  dimension; see remark 1.2.10) and exploited numerically in Chapter 2.

**Theorem 5.2.22 (maximum principle)** *If  $f \geq 0$  almost everywhere in  $\Omega$ , then  $u \geq 0$  almost everywhere in  $\Omega$ .*

**Remark 5.2.23** The maximum principle does nothing but express a property which is perfectly natural from the point of view of physics: for example, in the context of the stationary heat flow equation, if we have a heat source ( $f \geq 0$ ), the interior temperature is always greater than the boundary temperature ( $u \geq 0$ ). The maximum principle remains true if we replace the Laplacian by the more general operator (5.38) with variable coefficients in  $L^\infty(\Omega)$  or even if we consider the problem (5.16) with Neumann boundary conditions. The validity of the maximum principle is fundamentally linked to the ‘scalar’ character of the equation (that is, the unknown  $u$  has values in  $\mathbb{R}$ ). This maximum principle will generally fail if the unknown  $u$  is vector valued (for example, for the elasticity system (5.56)). •

**Proof.** We use the variational formulation (5.5) of (5.40) with  $v = u^- = \min(u, 0)$  which belongs to  $H_0^1(\Omega)$  from lemma 5.2.24 (since  $u = u^+ + u^-$ ). We have

$$\int_{\Omega} f u^- dx = \int_{\Omega} \nabla u \cdot \nabla u^- dx = \int_{\Omega} 1_{u < 0} \nabla u \cdot \nabla u dx = \int_{\Omega} |\nabla u^-|^2 dx \geq 0. \quad (5.41)$$

But  $u^- \leq 0$  and  $f \geq 0$  almost everywhere in  $\Omega$ . Consequently, all the terms of (5.41) are zero, and as  $u^- \in H_0^1(\Omega)$  we deduce that  $u^- = 0$ , that is,  $u \geq 0$  almost everywhere in  $\Omega$ . □

**Lemma 5.2.24** *If  $v \in H_0^1(\Omega)$ , then  $v^+ = \max(v, 0)$  belongs to  $H_0^1(\Omega)$  and*

$$\nabla v^+ = 1_{v > 0} \nabla v \text{ almost everywhere in } \Omega,$$

*where  $1_{v > 0}(x)$  is the function which equals 1 where  $v(x) > 0$  and 0 elsewhere.*

**Remark 5.2.25** The proof of lemma 5.2.24 is long and technical (it can be omitted in the first reading). We explain, nevertheless, why this proof is not simple, and in particular why we cannot use lemma 4.3.19 which says that a function defined ‘piecewise’, which belongs to  $H^1$  in each subdomain and is continuous in the interfaces between the subdomains, does actually belong to  $H^1$  in the whole domain. In effect, we would apply lemma 4.3.19 to  $v$  on the subdomain  $v > 0$  and to 0 on the subdomain  $v < 0$ , and we should be done. The problem is that, in general, the boundary of these two subdomains (defined by  $v = 0$ ) is not regular (in the sense of definition 3.2.5) even if  $v$  is a regular function. •



**Proof.** We show first that, if  $v \in H_0^1(\Omega)$  and if  $G(t)$  is a function from  $\mathbb{R}$  into  $\mathbb{R}$ , of class  $C^1$  such that  $G(0) = 0$  and  $G'(t)$  is bounded over  $\mathbb{R}$ , then  $G(v) \in H_0^1(\Omega)$  and  $\nabla(G(v)) = G'(v)\nabla v$ . By definition of  $H_0^1(\Omega)$ , there exists a sequence of functions  $v_n \in C_c^\infty(\Omega)$  which converges to  $v$  in the norm of  $H^1(\Omega)$  (in particular, for a subsequence,  $v_n$  and  $\nabla v_n$  converge almost everywhere in  $\Omega$  to  $v$  and  $\nabla v$  respectively ; see corollary 3.3.3 of [4]). We have

$$|G(v_n) - G(v)| \leq \left( \sup_{t \in \mathbb{R}} |G'(t)| \right) |v_n - v|, \quad (5.42)$$

therefore  $G(v_n)$  converge to  $G(v)$  in  $L^2(\Omega)$ . On the other hand, for  $1 \leq i \leq N$ ,

$$\left| \frac{\partial G(v_n)}{\partial x_i} - G'(v) \frac{\partial v}{\partial x_i} \right| \leq |G'(v_n) - G'(v)| \left| \frac{\partial v}{\partial x_i} \right| + \left( \sup_{t \in \mathbb{R}} |G'(t)| \right) \left| \frac{\partial v_n}{\partial x_i} - \frac{\partial v}{\partial x_i} \right|. \quad (5.43)$$

As  $|G'(v_n) - G'(v)| \left| \frac{\partial v}{\partial x_i} \right|$  converges almost everywhere to 0 (for a subsequence) and is bounded above by  $2 \left( \sup |G'(t)| \right) \left| \frac{\partial v}{\partial x_i} \right|$  which belongs to  $L^2(\Omega)$ , by application of the Lebesgue dominated convergence theorem (see [4], [35], [38]) this sequence of functions converges to 0 in  $L^2(\Omega)$ . The last term in (5.43) also converges to 0 in  $L^2(\Omega)$ , therefore  $G(v_n)$  is a Cauchy sequence in  $H_0^1(\Omega)$  which is a Hilbert space: it converges to a limit  $w \in H_0^1(\Omega)$ . By identification of the limits, we find that  $w = G(v)$  in  $L^2(\Omega)$  and that  $\frac{\partial w}{\partial x_i} = G'(v) \frac{\partial v}{\partial x_i}$  in  $L^2(\Omega)$ , from which we have the result.

We shall now approximate the function  $t \rightarrow \max(t, 0)$  by a sequence of functions  $G_n(t)$  of the type above to show that  $v^+$  belongs to  $H_0^1(\Omega)$ . Let  $G(t)$  be a function of  $C^1(\mathbb{R})$  such that

$$G(t) = 0 \text{ if } t \leq \frac{1}{2}, \quad 0 \leq G'(t) \leq 1 \text{ if } \frac{1}{2} \leq t \leq 1, \quad G'(t) = 1 \text{ if } 1 \leq t.$$

We define  $G_n(t) = G(nt)/n$  for  $n \geq 1$ , and we know by the arguments above that  $G_n(v) \in H_0^1(\Omega)$  and  $\frac{\partial G_n(v)}{\partial x_i} = G'_n(v) \frac{\partial v}{\partial x_i}$ . On the other hand, we are satisfied that

$$|G_n(v) - v^+| \leq \sup_{t \in \mathbb{R}} |G_n(t) - t^+| \leq \frac{1}{n},$$

therefore  $G_n(v)$  converges to  $v^+$  in  $L^2(\Omega)$ . We have also, for all  $1 \leq i \leq N$ ,

$$\left| \frac{\partial G_n(v)}{\partial x_i} - 1_{v>0} \frac{\partial v}{\partial x_i} \right| = |G'_n(v) - 1_{v>0}| \left| \frac{\partial v}{\partial x_i} \right| \leq 1_{0 < v < 1/n} \left| \frac{\partial v}{\partial x_i} \right|,$$

and as  $1_{0 < v < 1/n}$  converges to 0 almost everywhere, the Lebesgue dominated convergence theorem proves that  $\frac{\partial G_n(v)}{\partial x_i}$  converges to  $1_{v>0} \frac{\partial v}{\partial x_i}$  in  $L^2(\Omega)$ . We then deduce, as before, that  $G_n(v)$  converges to  $v^+$  in  $H_0^1(\Omega)$  and that  $\nabla v^+ = 1_{v>0} \nabla v$ .  $\square$

**Exercise 5.2.13** Show that the (nonlinear) mapping  $v \rightarrow v^+$  is continuous from  $L^2(\Omega)$  into itself, and also from  $H^1(\Omega)$  into itself (use the fact that  $\nabla u = 0$  almost everywhere on the set  $u^{-1}(0)$ ).

## Regularity

We now show that the solution of an elliptic boundary value problem is more regular than proved if the data is more regular than necessary.

**Theorem 5.2.26 (regularity)** *Take an integer  $m \geq 0$ . Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$  of class  $\mathcal{C}^{m+2}$ . Let  $f \in H^m(\Omega)$ . Then, the unique solution  $u \in H_0^1(\Omega)$  of (5.40) belongs to  $H^{m+2}(\Omega)$ . Further, the mapping  $f \rightarrow u$  is linear and continuous from  $H^m(\Omega)$  into  $H^{m+2}(\Omega)$ , that is, there exists a constant  $C > 0$  such that*

$$\|u\|_{H^{m+2}(\Omega)} \leq C \|f\|_{H^m(\Omega)}.$$

By immediate application of the regularity theorem 5.2.26 and of the theorem 4.3.25 (on the continuity of functions of  $H^m(\Omega)$ ), we obtain as a corollary the result stated before in remark 5.2.3, that the weak solutions of elliptic PDEs are in fact strong (or classical) solutions if the data is regular.

**Corollary 5.2.27** *If  $\Omega$  is an open bounded set of  $\mathbb{R}^N$  of class  $\mathcal{C}^{m+2}$ , if  $f \in H^m(\Omega)$ , and if  $m > N/2$ , then the variational solution  $u \in H_0^1(\Omega)$  of (5.40) is a strong solution since it belongs to  $C^2(\bar{\Omega})$ .*

*In particular, if  $\Omega$  is an open bounded set of  $\mathbb{R}^N$  of class  $\mathcal{C}^\infty$ , and if  $f \in C^\infty(\bar{\Omega})$ , then the solution  $u \in H_0^1(\Omega)$  of (5.40) is also in  $C^\infty(\bar{\Omega})$ .*

**Remark 5.2.28** It is important to understand the effect of these regularity results. Assuming that  $\Delta u$ , which is a particular combination of some second derivatives of  $u$ , belongs to a certain function space, we deduce that **all** the second derivatives of  $u$  belong to this same space! Of course, all these regularity results are obvious in  $N = 1$  dimensions since the Laplacian coincides with the second derivative and therefore the equation directly gives the regularity of this second derivative. •

**Remark 5.2.29** The regularity theorem 5.2.26 and its corollary 5.2.27 remain valid for Neumann boundary conditions. It generalizes also to the case of elliptic operators with variable coefficients (as in Section 5.2.3). In this last case, we must add to the usual hypothesis of coercivity of the coefficients, the hypothesis that the coefficients are of class  $\mathcal{C}^{m+1}$  in  $\Omega$  (when  $\Omega$  is bounded regular of class  $\mathcal{C}^{m+2}$  and that  $f \in H^m(\Omega)$ ). •

We shall not prove the regularity theorem 5.2.26 in all its generality but only in a particular case which is more simple (we shall explain in remark 5.2.32 how to pass from the particular case to the general case). We take  $\Omega = \mathbb{R}^N$ , and for  $f \in L^2(\mathbb{R}^N)$  we consider the problem

$$-\Delta u + u = f \quad \text{in } \mathbb{R}^N. \quad (5.44)$$

There are no **explicit** boundary conditions in (5.44) as there is no boundary. Nevertheless, the behaviour at infinity of the solution is a kind of boundary condition. By taking  $f$  in  $L^2(\mathbb{R}^N)$  we have chosen an **implicit** boundary condition which is to look for  $u$  in  $H^1(\mathbb{R}^N)$ ,

that is, in a certain sense,  $u(x)$  ‘tends to zero’ at infinity so that the integral  $\int_{\mathbb{R}^N} |u|^2 dx$  converges. A variational formulation of (5.44) is: find  $u \in H^1(\mathbb{R}^N)$  such that

$$\int_{\mathbb{R}^N} (\nabla u \cdot \nabla v + uv) dx = \int_{\mathbb{R}^N} f v dx \quad \forall v \in H^1(\mathbb{R}^N). \quad (5.45)$$

A direct application of the Lax–Milgram theorem 3.3.1 proves that there exists a unique solution  $u \in H^1(\mathbb{R}^N)$  of (5.45). Finally, by identical reasoning to that already done in this chapter, show that this solution of the variational formulation is also a solution of the PDE (5.44). We can now state the regularity result.

**Proposition 5.2.30** *If  $f \in L^2(\mathbb{R}^N)$ , then the solution  $u \in H^1(\mathbb{R}^N)$  of (5.44) belongs to  $H^2(\mathbb{R}^N)$ . Likewise, if  $f \in H^m(\mathbb{R}^N)$  (with  $m \geq 0$ ), then  $u$  belongs to  $H^{m+2}(\mathbb{R}^N)$ .*

**Proof.** The essential ingredient of the proof is the ‘method of translation’. For  $h \in \mathbb{R}^N$ ,  $h \neq 0$ , we define a difference quotient

$$D_h v(x) = \frac{v(x+h) - v(x)}{|h|}$$

which belongs to  $H^1(\mathbb{R}^N)$  if  $v \in H^1(\mathbb{R}^N)$ . We easily see that  $\nabla(D_h v) = D_h(\nabla v)$  and that, for  $v, \phi \in L^2(\mathbb{R}^N)$ , we have a ‘discrete integration by parts formula’

$$\int_{\mathbb{R}^N} (D_h v) \phi dx = \int_{\mathbb{R}^N} v (D_{-h} \phi) dx.$$

Other properties of the quotient  $D_h v$  are given in lemma 5.2.31. In the variational formulation (5.45) we take  $v = D_{-h}(D_h u)$ , and applying the rules above we obtain,

$$\int_{\mathbb{R}^N} (|\nabla(D_h u)|^2 + |D_h u|^2) dx = \int_{\mathbb{R}^N} f D_{-h}(D_h u) dx.$$

We deduce the upper bound

$$\|D_h u\|_{H^1(\mathbb{R}^N)}^2 \leq \|f\|_{L^2(\mathbb{R}^N)} \|D_{-h}(D_h u)\|_{L^2(\mathbb{R}^N)}.$$

Now, by application of (5.48), we have also

$$\|D_{-h}(D_h u)\|_{L^2(\mathbb{R}^N)} \leq \|\nabla(D_h u)\|_{L^2(\mathbb{R}^N)} \leq \|D_h u\|_{H^1(\mathbb{R}^N)}.$$

Therefore, we have  $\|D_h u\|_{H^1(\mathbb{R}^N)} \leq \|f\|_{L^2(\mathbb{R}^N)}$ , and in particular, for  $1 \leq i \leq N$ ,

$$\left\| D_h \frac{\partial u}{\partial x_i} \right\|_{L^2(\mathbb{R}^N)} \leq \|f\|_{L^2(\mathbb{R}^N)},$$

which implies, from lemma 5.2.31 below, that  $\frac{\partial u}{\partial x_i}$  belongs to  $H^1(\mathbb{R}^N)$ , that is,  $u \in H^2(\mathbb{R}^N)$ .

Suppose now that  $f \in H^1(\mathbb{R}^N)$ . We show that  $\frac{\partial u}{\partial x_i}$  is the unique solution in  $H^1(\mathbb{R}^N)$  of

$$-\Delta u_i + u_i = \frac{\partial f}{\partial x_i} \quad \text{in } \mathbb{R}^N. \quad (5.46)$$

If this is true, by application of the preceding part of the proof, we deduce that  $\frac{\partial u}{\partial x_i}$  belongs to  $H^2(\mathbb{R}^N)$ , that is,  $u \in H^3(\mathbb{R}^N)$  as seen. We write the variational formulation (5.45) with the test function  $\frac{\partial \phi}{\partial x_i}$  for  $\phi \in C_c^\infty(\mathbb{R}^N)$

$$\int_{\mathbb{R}^N} \left( \nabla u \cdot \nabla \frac{\partial \phi}{\partial x_i} + u \frac{\partial \phi}{\partial x_i} \right) dx = \int_{\mathbb{R}^N} f \frac{\partial \phi}{\partial x_i} dx.$$

As  $u \in H^2(\mathbb{R}^N)$  and  $f \in H^1(\mathbb{R}^N)$ , we can integrate by parts to obtain

$$\int_{\mathbb{R}^N} \left( \nabla \frac{\partial u}{\partial x_i} \cdot \nabla \phi + \frac{\partial u}{\partial x_i} \phi \right) dx = \int_{\mathbb{R}^N} \frac{\partial f}{\partial x_i} \phi dx \quad (5.47)$$

which, by density, remains true for all  $\phi \in H^1(\mathbb{R}^N)$ . We establish that (5.47) is the variational formulation of (5.46). Consequently,  $\frac{\partial u}{\partial x_i} = u_i$  is the unique solution in  $H^1(\mathbb{R}^N)$  of (5.46).

The case  $f \in H^m(\mathbb{R}^N) \Rightarrow u \in H^{m+2}(\mathbb{R}^N)$  is proved by induction on  $m$  as we have done for  $m = 1$ .  $\square$

**Lemma 5.2.31** *For  $v \in L^2(\mathbb{R}^N)$ ,  $h \in \mathbb{R}^N$ ,  $h \neq 0$ , we define a difference quotient*

$$D_h v(x) = \frac{v(x+h) - v(x)}{|h|} \in L^2(\mathbb{R}^N).$$

*If  $v \in H^1(\mathbb{R}^N)$ , we have the estimate*

$$\|D_h v\|_{L^2(\mathbb{R}^N)} \leq \|\nabla v\|_{L^2(\mathbb{R}^N)}. \quad (5.48)$$

*Conversely, let  $v \in L^2(\mathbb{R}^N)$ : if there exists a constant  $C$ , such that, for all  $h \neq 0$ , we have*

$$\|D_h v\|_{L^2(\mathbb{R}^N)} \leq C, \quad (5.49)$$

*then  $v \in H^1(\mathbb{R}^N)$  and  $\|e \cdot \nabla v\|_{L^2(\mathbb{R}^N)} \leq C$  for every unit vector  $e \in \mathbb{R}^N$ .*

**Proof.** We prove (5.48) first of all. For  $v \in C_c^\infty(\mathbb{R}^N)$ , we write

$$D_h v(x) = \int_0^1 \frac{h}{|h|} \cdot \nabla v(x+th) dt,$$

which is bounded above by

$$|D_h v(x)|^2 \leq \int_0^1 |\nabla v(x+th)|^2 dt.$$

Integrating in  $x$  it becomes

$$\|D_h v\|_{L^2(\mathbb{R}^N)}^2 \leq \int_0^1 \int_{\mathbb{R}^N} |\nabla v(x+th)|^2 dx dt \leq \int_0^1 \|\nabla v\|_{L^2(\mathbb{R}^N)}^2 dt = \|\nabla v\|_{L^2(\mathbb{R}^N)}^2.$$

By density, the estimate (5.48) is true for every function of  $H^1(\mathbb{R}^N)$ .

Now let  $v \in L^2(\mathbb{R}^N)$  which satisfies the estimate (5.49). We deduce that, for all  $\phi \in C_c^\infty(\mathbb{R}^N)$ ,

$$\left| \int_{\mathbb{R}^N} D_h v \phi dx \right| \leq C \|\phi\|_{L^2(\mathbb{R}^N)}.$$

Now, by discrete integration by parts, we have

$$\int_{\mathbb{R}^N} (D_h v) \phi \, dx = \int_{\mathbb{R}^N} v (D_{-h} \phi) \, dx,$$

and, since  $\phi$  is regular, if we set  $h = te$  with  $e \in \mathbb{R}^N$ ,  $e \neq 0$ , we have

$$\lim_{t \rightarrow 0} D_{-h} \phi(x) = -e \cdot \nabla \phi(x).$$

Consequently, we deduce that, for  $1 \leq i \leq N$  and all  $\phi \in C_c^\infty(\mathbb{R}^N)$ , we have

$$\left| \int_{\mathbb{R}^N} v \frac{\partial \phi}{\partial x_i} \, dx \right| \leq C \|\phi\|_{L^2(\mathbb{R}^N)},$$

which is nothing other than the definition of  $v$  belonging to  $H^1(\mathbb{R}^N)$  (see definition 4.3.1).  $\square$

**Remark 5.2.32 (delicate)** Let us now explain how we pass from the case treated in proposition 5.2.30 (with  $\Omega = \mathbb{R}^N$ ) to the general case of theorem 5.2.26. We use an argument of local coordinates and a partition of unity as in the proof of proposition 4.4.2. Following the notation of definition 3.2.5 of a regular open set (see also Figure 4.4), there exists a finite covering of  $\Omega$  by open sets  $(\omega_i)_{0 \leq i \leq I}$  and a partition of unity  $(\theta_i)_{0 \leq i \leq I}$  such that

$$\theta_i \in C_c^\infty(\omega_i), \quad 0 \leq \theta_i(x) \leq 1, \quad \sum_{i=0}^I \theta_i(x) = 1 \quad \text{in } \overline{\Omega}.$$

The open set  $\omega_0$  is in the interior of  $\Omega$  (in fact  $\overline{\omega_0} \subset \Omega$ ) while the other open sets  $\omega_i$ , for  $i \geq 1$ , cover the boundary  $\partial\Omega$ . Let  $u \in H_0^1(\Omega)$  be the unique solution of (5.40). To show the regularity of  $u = \sum_{i=0}^I \theta_i u$ , we shall show the regularity of each of the terms  $\theta_i u$ . For the term  $\theta_0 u$  we talk of the interior regularity of  $u$ . It is immediate that  $\theta_0 u$  is the unique solution in  $H^1(\mathbb{R}^N)$  of

$$-\Delta(\theta_0 u) + \theta_0 u = f_0 \quad \text{in } \mathbb{R}^N,$$

with  $f_0 = \theta_0(f - u) - 2\nabla\theta_0 \cdot \nabla u - u\Delta\theta_0$  which belongs to  $L^2(\mathbb{R}^N)$ . By application of proposition 5.2.30 we deduce therefore that  $\theta_0 u \in H^2(\mathbb{R}^N)$ . This allows us to improve the regularity of  $f_0$ , and by successive application of proposition 5.2.30 we conclude that  $f \in H^m(\Omega)$  implies that  $\theta_0 u \in H^{m+2}(\Omega)$ . To show the regularity of the other terms  $\theta_i u$  for  $i \geq 1$ , we must initially ‘rectify’ the boundary to reduce it to the case  $\Omega = \mathbb{R}_+^N$ . We must therefore show a result of the same type as proposition 5.2.30 but for  $\Omega = \mathbb{R}_+^N$ . This is a little more delicate because when rectifying the boundary by local coordinates we have to change the coefficients of the elliptic operator (the Laplacian becomes an operator with variable coefficients as in Section 5.2.3): we refer to [6] for the details. To summarize, it is enough that we have the regularity in all the space and in a half-space to prove the regularity in a regular open bounded set.  $\bullet$

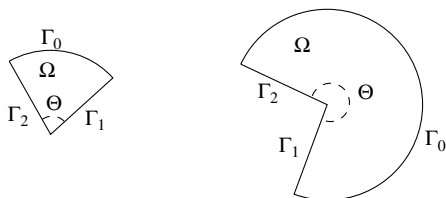


Figure 5.2. Angular sector  $\Omega$  with angle  $\Theta$  (smaller or larger than  $\pi$ ).

### Example of singularity

Let us now see an example of **singular** solutions, that is, solutions which are nonregular. We consider a problem posed in an open nonregular set for which the regularity theorem 5.2.26 and its corollary 5.2.27 are false. In particular, even if weak solutions exist (that is, belonging to the Sobolev space  $H^1(\Omega)$ ) they are not strong solutions (that is, twice differentiable) for this problem. We work in  $N = 2$  space dimensions, and we consider an angular sector  $\Omega$  defined in radial coordinates by (see Figure 5.2)

$$\Omega = \{(r, \theta) \text{ such that } 0 \leq r < R \text{ and } 0 < \theta < \Theta\}$$

with  $0 < R < +\infty$  and  $0 < \Theta \leq 2\pi$  (recall that  $x_1 = r \cos \theta$  and  $x_2 = r \sin \theta$ ). We denote by  $\Gamma_0$  the part of the boundary of  $\Omega$  where  $r = R$ ,  $\Gamma_1$  where  $\theta = 0$ , and  $\Gamma_2$  where  $\theta = \Theta$ . This open set  $\Omega$  has three ‘corners’, but only the origin (the corner between the boundaries  $\Gamma_1$  and  $\Gamma_2$ ) can cause problems for the regularity in the examples below. Physically, the case of an angle  $\Theta < \pi$  is representative of a tip effect, while the case  $\Theta > \pi$  corresponds to a notch (or a **fissure** if  $\Theta = 2\pi$ ).

For an integer  $k \geq 1$ , we study the following two boundary value problems

$$\begin{cases} -\Delta u = 0 & \text{in } \Omega \\ u = \cos\left(\frac{k\pi\theta}{\Theta}\right) & \text{on } \Gamma_0 \\ \frac{\partial u}{\partial n} = 0 & \text{on } \Gamma_1 \cup \Gamma_2 \end{cases} \quad (5.50)$$

and

$$\begin{cases} -\Delta u = 0 & \text{in } \Omega \\ u = \sin\left(\frac{k\pi\theta}{\Theta}\right) & \text{on } \Gamma_0 \\ u = 0 & \text{on } \Gamma_1 \cup \Gamma_2 \end{cases} \quad (5.51)$$

We could study the regularity of the solutions of (5.50) and (5.51) in terms of whether or not they belong to the Sobolev spaces  $H^m(\Omega)$ , but, for simplicity and to keep an obvious physical meaning in our results, we shall simply be interested in the behaviour of the gradient  $\nabla u$  in a neighbourhood of the origin. From the viewpoint of physics or mechanics, this gradient corresponds to a heat flux, an electric field, or a stress field: it is important to know if this quantity is continuously bounded or not at the origin.

**Lemma 5.2.33** *There exists a unique weak solution of (5.50) in  $H^1(\Omega)$ , given by the formula*

$$u(r, \theta) = \left(\frac{r}{R}\right)^{k\pi/\Theta} \cos\left(\frac{k\pi\theta}{\Theta}\right). \quad (5.52)$$

*Likewise, there exists a unique weak solution of (5.51) in  $H^1(\Omega)$ , given by the formula*

$$u(r, \theta) = \left(\frac{r}{R}\right)^{k\pi/\Theta} \sin\left(\frac{k\pi\theta}{\Theta}\right). \quad (5.53)$$

*In the two cases, if  $k = 1$  and  $\pi < \Theta$ , then the gradient  $\nabla u$  is not bounded at the origin, while if  $k \geq 2$  or  $\pi \geq \Theta$ , then the gradient  $\nabla u$  is continuous at the origin.*

**Remark 5.2.34** When we impose more general Dirichlet data in problems (5.50) and (5.51), we can decompose them into a Fourier series in  $\theta$  on  $\Gamma_0$  and apply lemma 5.2.33 to each term of the series (nevertheless, this Dirichlet data on  $\Gamma_0$  must be compatible with the boundary conditions on  $\Gamma_1$  and  $\Gamma_2$ ). We deduce that, if  $\Theta \leq \pi$ , then the solutions of (5.50) and (5.51) are always regular no matter what the Dirichlet data on  $\Gamma_0$ . Conversely, if  $\Theta > \pi$ , the solutions of (5.50) and (5.51) may be singular.

Physically, we interpret this regularity result by saying that a notch ( $\Theta > \pi$ ) causes a singularity, as opposed to a point ( $\Theta \leq \pi$ ). In this case, it is sometimes necessary to re-examine the modelling since a heat flux, electric field, or a stress field which is infinite at the origin does not have a physical sense. Two approximations in the model may be at the origin of this nonphysical singularity: on the one hand, an angle is never ‘perfect’ but often a little ‘rounded’, on the other hand, when  $\nabla u$  is very large, we leave the domain of validity of the linear equations that we are studying (typically, a constitutive law such as Fourier’s law (1.3) becomes nonlinear as the thermal conductivity is itself a function of  $\nabla u$ ). In any event, a regularity result gives important results on the limits of application of the model.

In the case where  $\Theta = 2\pi$ , the open set  $\Omega$  is a **cracked** domain and lemma 5.2.33 has a very important mechanical interpretation if problems (5.50) and (5.51) model the antiplane shearing of a cylinder with base  $\Omega$  (see exercise 5.3.7). In this case, the coefficient of the term of order  $k = 1$  in the Fourier series (which leads to the singular behaviour at the origin) is called the **stress intensity factor** which is often used in models of crack propagation or rupture (see, for example, [26]). •

**Remark 5.2.35** The singular solutions given by lemma 5.2.33 are not only theoretical counterexamples: they can be shown numerically (see Figure 6.18). We also add that these singular solutions are a source of difficulty for numerical methods (see remark 6.3.15) which confirms the interest in their study. •

**Remark 5.2.36** We can also consider the case of Dirichlet boundary conditions on  $\Gamma_1$  and Neumann on  $\Gamma_2$  (in this case, we must take  $u = \sin(\frac{k\pi\theta}{2\Theta})$  on  $\Gamma_0$ ). The only difference relates to the case  $k\pi = \Theta$  for which the gradient  $\nabla u$  is bounded but not continuous at the origin. In particular, we find that for  $k = 1$  and  $\pi = \Theta$ , **although**

**there is no corner**, the solution is singular. This is due to the change of boundary conditions on a regular part of the boundary. •

**Proof.** We limit ourselves to treating the problem (5.50) with Neumann boundary conditions in the corner (the other problem (5.51) is treated exactly in the same way). We start by ‘lifting’ the nonhomogeneous boundary conditions by defining a function  $u_0 \in H^1(\Omega)$  whose trace on the boundary coincides with these boundary conditions. We can easily check that

$$u_0(r, \theta) = \frac{r^2}{R^2} \cos\left(\frac{k\pi\theta}{\Theta}\right)$$

is satisfactory, that is,  $u = u_0 + v$  where  $v$  is the solution of a homogeneous problem

$$\begin{cases} -\Delta v = \Delta u_0 & \text{in } \Omega \\ v = 0 & \text{on } \Gamma_0 \\ \frac{\partial v}{\partial n} = 0 & \text{on } \Gamma_1 \cup \Gamma_2. \end{cases} \quad (5.54)$$

We remark that this lifting is possible since the data on  $\Gamma_0$  is compatible with the boundary conditions on  $\Gamma_1$  and  $\Gamma_2$ , that is, in this case its derivative in  $\theta$  (that is, its normal derivative) is zero for  $\theta = 0$  or  $\Theta$ . As  $\Delta u_0$  belongs to  $L^2(\Omega)$ , there exists a unique solution of (5.54) in  $H^1(\Omega)$ , and consequently (5.50) also has a unique solution in  $H^1(\Omega)$ . We verify that (5.52) is precisely this unique solution. Recall that

$$\Delta\phi(r, \theta) = \frac{\partial^2\phi}{\partial r^2} + \frac{1}{r} \frac{\partial\phi}{\partial r} + \frac{1}{r^2} \frac{\partial^2\phi}{\partial\theta^2}.$$

A simple calculation shows that

$$\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r}\right) \left(\frac{r}{R}\right)^{k\pi/\Theta} = \left(\frac{k\pi}{\Theta}\right)^2 \frac{1}{r^2} \left(\frac{r}{R}\right)^{k\pi/\Theta},$$

therefore (5.52) is also a solution of (5.50). We also see easily that (5.52) belongs to  $H^1(\Omega)$ . Finally, the formula for the gradient in radial coordinates in the basis  $(e_r, e_\theta)$

$$\nabla\phi(r, \theta) = \frac{\partial\phi}{\partial r} e_r + \frac{1}{r} \frac{\partial\phi}{\partial\theta} e_\theta$$

gives us

$$\nabla u = \left(\frac{r}{R}\right)^{(k\pi/\Theta)-1} \frac{k\pi}{\Theta} \left( \cos\left(\frac{k\pi\theta}{\Theta}\right) e_r - \sin\left(\frac{k\pi\theta}{\Theta}\right) e_\theta \right)$$

which, clearly, is not bounded at the origin if  $0 < k\pi < \Theta$  and is continuous at the origin if  $k\pi > \Theta$ . The limiting case  $k\pi = \Theta$  also corresponds to a continuous gradient since in fact  $u = x_1$ !  $\square$



## 5.3 Solution of other models

### 5.3.1 System of linearized elasticity

We apply the variational approach to the solution of the system of linearized elasticity equations. We start by describing the mechanical model which we have seen in a particular case in Chapter 1. These equations model the deformations of a solid under the hypothesis of small deformations and small displacements (this hypothesis allows us to obtain linear equations; from which we have the name **linear** elasticity, see for example, [36]). We consider the stationary elasticity equations, that is, independent of time. Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$ . Let a force  $f(x)$  be a function from  $\Omega$  into  $\mathbb{R}^N$ . The unknown  $u$  (the displacement) is also a function from  $\Omega$  into  $\mathbb{R}^N$ . The mechanical modelling uses the deformation tensor, denoted by  $e(u)$ , which is a function with values in the set of symmetric matrices

$$e(u) = \frac{1}{2} \left( \nabla u + (\nabla u)^t \right) = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)_{1 \leq i, j \leq N},$$

as well as the stress tensor  $\sigma$  (another function with values in the set of symmetric matrices) which is related to  $e(u)$  by Hooke's law

$$\sigma = 2\mu e(u) + \lambda \operatorname{tr}(e(u)) \mathbf{I},$$

where  $\lambda$  and  $\mu$  are the Lamé coefficients of the homogeneous isotropic material which occupies  $\Omega$ . For thermodynamic reasons the Lamé coefficients satisfy

$$\mu > 0 \quad \text{and} \quad 2\mu + N\lambda > 0.$$

We add, to this constitutive law, the balance of forces in the solid

$$-\operatorname{div} \sigma = f \text{ in } \Omega$$

where, by definition, the divergence of  $\sigma$  is the vector of components

$$\operatorname{div} \sigma = \left( \sum_{j=1}^N \frac{\partial \sigma_{ij}}{\partial x_j} \right)_{1 \leq i \leq N}.$$

Using the fact that  $\operatorname{tr}(e(u)) = \operatorname{div} u$ , we deduce the equations for  $1 \leq i \leq N$

$$-\sum_{j=1}^N \frac{\partial}{\partial x_j} \left( \mu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) + \lambda (\operatorname{div} u) \delta_{ij} \right) = f_i \text{ in } \Omega \quad (5.55)$$

with  $f_i$  and  $u_i$ , for  $1 \leq i \leq N$ , the components of  $f$  and  $u$  in the canonical basis of  $\mathbb{R}^N$ . By adding a Dirichlet boundary condition, and by using vector notation, the boundary value problem is

$$\begin{cases} -\operatorname{div} (2\mu e(u) + \lambda \operatorname{tr}(e(u)) \mathbf{I}) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (5.56)$$

We can state and prove a first existence and uniqueness result for the system of linear elasticity.

**Theorem 5.3.1** *Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$ . Let  $f \in L^2(\Omega)^N$ . There exists a unique (weak) solution  $u \in H_0^1(\Omega)^N$  of (5.56).*

**Proof.** To find the variational formulation we multiply each equation (5.55) by a test function  $v_i$  (which is zero on the boundary  $\partial\Omega$  to take account of the Dirichlet boundary conditions) and we integrate by parts to obtain

$$\int_{\Omega} \mu \sum_{j=1}^N \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \frac{\partial v_i}{\partial x_j} dx + \int_{\Omega} \lambda \operatorname{div} u \frac{\partial v_i}{\partial x_i} dx = \int_{\Omega} f_i v_i dx.$$

We sum these equations, for  $i$  going from 1 to  $N$ , in order to obtain the divergence of the function  $v = (v_1, \dots, v_N)$  and to simplify the first integral as

$$\sum_{i,j=1}^N \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \frac{\partial v_i}{\partial x_j} = \frac{1}{2} \sum_{i,j=1}^N \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) = 2e(u) \cdot e(v).$$

Choosing  $H_0^1(\Omega)^N$  as the Hilbert space, we obtain the variational formulation: find  $u \in H_0^1(\Omega)^N$  such that

$$\int_{\Omega} 2\mu e(u) \cdot e(v) dx + \int_{\Omega} \lambda \operatorname{div} u \operatorname{div} v dx = \int_{\Omega} f \cdot v dx \quad \forall v \in H_0^1(\Omega)^N. \quad (5.57)$$

We easily see that each term of (5.57) has a meaning.

To be able to apply the Lax–Milgram theorem 3.3.1 to the variational formulation (5.57), the only delicate hypothesis to verify is the coercivity of the bilinear form. We proceed in three stages. First, we show that

$$\int_{\Omega} 2\mu |e(v)|^2 dx + \int_{\Omega} \lambda |\operatorname{div} v|^2 dx \geq \nu \int_{\Omega} |e(v)|^2 dx,$$

with  $\nu = \min(2\mu, (2\mu + N\lambda)) > 0$ . For this, we use an algebraic inequality: if we denote by  $A \cdot B = \sum_{i,j=1}^N a_{ij} b_{ij}$  the usual scalar product of symmetric matrices, we can decompose every real symmetric matrix  $A$  in the form

$$A = A^d + A^h \text{ with } A^d = A - \frac{1}{N} \operatorname{tr} A \mathbf{I} \text{ and } A^h = \frac{1}{N} \operatorname{tr} A \mathbf{I},$$

in such a way that  $A^d \cdot A^h = 0$  and  $|A|^2 = |A^d|^2 + |A^h|^2$ . We then have

$$2\mu |A|^2 + \lambda (\operatorname{tr} A)^2 = 2\mu |A^d|^2 + (2\mu + N\lambda) |A^h|^2 \geq \nu |A|^2$$

with  $\nu = \min(2\mu, (2\mu + N\lambda))$ , which gives the result for  $A = e(u)$ . The fact that we assume  $\nu > 0$  is not a problem: the mechanical and thermodynamical arguments

which lead to the inequalities  $\mu > 0$  and  $(2\mu + N\lambda) > 0$  are exactly the same. Second, we use Korn's inequality (or a simple case of this inequality, see lemma 5.3.2) which gives a constant  $C > 0$  such that

$$\int_{\Omega} |e(v)|^2 dx \geq C \int_{\Omega} |\nabla v|^2 dx$$

for all  $v \in H_0^1(\Omega)^N$ . Third, we use the Poincaré inequality (component by component, see proposition 4.3.10) which gives a constant  $C > 0$  such that, for all  $v \in H_0^1(\Omega)^N$ ,

$$\int_{\Omega} |v|^2 dx \leq C \int_{\Omega} |\nabla v|^2 dx.$$

In summary, these three inequalities lead to the coercivity

$$\int_{\Omega} 2\mu |e(v)|^2 dx + \int_{\Omega} \lambda |\operatorname{div} v|^2 dx \geq C \|v\|_{H^1(\Omega)}^2.$$

The Lax–Milgram theorem 3.3.1 therefore gives the existence and uniqueness of the solution of the variational formulation (5.57). Finally, to show that the unique solution of (5.57) is also a solution of the boundary value problem (5.56), we proceed as in the proof of the theorem 5.2.2 for the Laplacian.  $\square$

**Lemma 5.3.2** *Let  $\Omega$  be an open set of  $\mathbb{R}^N$ . For every function  $v \in H_0^1(\Omega)^N$ , we have*

$$\|\nabla v\|_{L^2(\Omega)} \leq \sqrt{2} \|e(v)\|_{L^2(\Omega)}. \quad (5.58)$$

**Proof.** Take  $v \in C_c^\infty(\Omega)^N$ . By integration by parts we obtain

$$2 \int_{\Omega} |e(v)|^2 dx = \int_{\Omega} |\nabla v|^2 dx + \int_{\Omega} \nabla v \cdot (\nabla v)^t dx = \int_{\Omega} |\nabla v|^2 dx + \int_{\Omega} |\operatorname{div} v|^2 dx.$$

By density of  $C_c^\infty(\Omega)$  in  $H_0^1(\Omega)$ , we deduce (5.58).  $\square$

**Exercise 5.3.1** Show that the mapping of  $L^2(\Omega)^N$  into  $H_0^1(\Omega)^N$ , which maps each  $f$  into a  $u$ , the unique weak solution of (5.56), is linear and continuous.

The analysis of the boundary value problem (5.56) with Dirichlet boundary conditions **on all the boundary**  $\partial\Omega$  is a little misleading in its simplicity. In effect, to introduce other boundary conditions (for example, Neumann) on part of the boundary, the proof of the coercivity of the variational formulation becomes much more difficult since we must replace the elementary inequality of lemma 5.3.2 by its generalization, which is more technical, called Korn's inequality (see lemma 5.3.3 below). Recall that we cannot, in general, be content with a Dirichlet condition on the whole of the boundary  $\partial\Omega$  since it means that the solid is fixed and immobile on its boundary. In practice, all the boundary is not fixed and often a part of the boundary is free

to move, or surface forces are applied to another part. These two cases are modelled by Neumann boundary conditions which are written here

$$\sigma n = g \text{ on } \partial\Omega, \quad (5.59)$$

where  $g$  is a vector valued function. The Neumann condition (5.59) is interpreted by saying that  $g$  is a force applied on the boundary. If  $g = 0$ , we say that no force is applied and the boundary can move without restriction: we say that the boundary is free.

We shall now consider the elasticity system with mixed boundary conditions (a mixture of Dirichlet and of Neumann), that is,

$$\begin{cases} -\operatorname{div}(2\mu e(u) + \lambda \operatorname{tr}(e(u))I) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega_D \\ \sigma n = g & \text{on } \partial\Omega_N, \end{cases} \quad (5.60)$$

where  $(\partial\Omega_N, \partial\Omega_D)$  is a partition of  $\partial\Omega$  such that the surface measures of  $\partial\Omega_N$  and  $\partial\Omega_D$  are nonzero (see Figure 4.1). The analysis of this new boundary value problem is more complicated than in the case of Dirichlet boundary condition: we must use Korn's inequality below.

**Lemma 5.3.3 (Korn's inequality)** *Let  $\Omega$  be a regular open bounded set of class  $\mathcal{C}^1$  of  $\mathbb{R}^N$ . There exists a constant  $C > 0$  such that, for every function  $v \in H^1(\Omega)^N$ , we have*

$$\|v\|_{H^1(\Omega)} \leq C \left( \|v\|_{L^2(\Omega)}^2 + \|e(v)\|_{L^2(\Omega)}^2 \right)^{1/2}. \quad (5.61)$$

The inequality (5.61) is not banal: in effect, its left-hand side contains all the partial derivatives of  $v$  and its right-hand side involves only certain linear combinations of partial derivatives. As the inverse of (5.61) is obvious, we deduce that the two sides of (5.61) are equivalent norms. The proof of the lemma 5.3.3 is complicated and is outside the scope of this course (see, for example, [15]). We remark that we have proved Korn's inequality (5.61) when  $v$  belongs to  $H_0^1(\Omega)^N$ . In effect, in this case a combination of lemma 5.3.2 and of the Poincaré inequality (for an open bounded set) gives the inequality (5.61).

The mechanical interpretation of Korn's inequality is the following. The elastic energy, proportional to the norm of the deformation tensor  $e(u)$  in  $L^2(\Omega)$ , controls the norm of the displacement  $u$  in  $H^1(\Omega)^N$ , with the addition of the norm of  $u$  in  $L^2(\Omega)$ . As we shall see in exercise 5.3.2, this last addition is to take account of the **rigid body motion**, that is, the displacement  $u$  which is nonzero but with elastic energy zero.

**Exercise 5.3.2** Let  $\Omega$  be an open connected set of  $\mathbb{R}^N$ . Let the set  $\mathcal{R}$  be the 'rigid motions' of  $\Omega$  defined by

$$\mathcal{R} = \{v(x) = b + Mx \text{ with } b \in \mathbb{R}^N, M = -M^t \text{ an antisymmetric matrix} \}. \quad (5.62)$$

Show that  $v \in H^1(\Omega)^N$  satisfies  $e(v) = 0$  in  $\Omega$  if and only if  $v \in \mathcal{R}$ .

We can now state a second existence and uniqueness result for the system of linear elasticity with mixed boundary conditions.

**Theorem 5.3.4** *Let  $\Omega$  be a regular open bounded connected set of class  $C^1$  of  $\mathbb{R}^N$ . Let  $f \in L^2(\Omega)^N$  and  $g \in L^2(\partial\Omega_N)^N$ . We define the space*

$$V = \{v \in H^1(\Omega)^N \text{ such that } v = 0 \text{ on } \partial\Omega_D\}. \quad (5.63)$$

*There exists a unique (weak) solution  $u \in V$  of (5.60) which depends linearly and continuously on the data  $f$  and  $g$ .*

**Proof.** The variational formulation of (5.60) is obtained as in the proof of the theorem 5.3.1. The space  $V$ , defined by (5.63), contains the Dirichlet boundary condition on  $\partial\Omega_D$  and is a Hilbert space as a closed subspace of  $H^1(\Omega)^N$  (by application of the trace theorem 4.3.13). We then obtain the variational formulation: find  $u \in V$  such that

$$\int_{\Omega} 2\mu e(u) \cdot e(v) \, dx + \int_{\Omega} \lambda \operatorname{div} u \operatorname{div} v \, dx = \int_{\Omega} f \cdot v \, dx + \int_{\partial\Omega_N} g \cdot v \, ds \quad \forall v \in V. \quad (5.64)$$

To be able to apply the Lax–Milgram theorem 3.3.1 to the variational formulation (5.64), the only delicate hypothesis to be verified is once again the coercivity of the bilinear form. In other words, we must show that there exists a constant  $C > 0$  such that, for every function  $v \in V$ , we have

$$\|v\|_{H^1(\Omega)} \leq C \|e(v)\|_{L^2(\Omega)}. \quad (5.65)$$

First, we note that  $\|e(v)\|_{L^2(\Omega)}$  is a norm over  $V$ . The only point to be verified which should delay us is that  $\|e(v)\|_{L^2(\Omega)} = 0$  implies that  $v = 0$ . Suppose therefore that  $\|e(v)\|_{L^2(\Omega)} = 0$ : then exercise 5.3.2 shows that  $v$  is a rigid displacement, that is,  $v(x) = b + Mx$  with  $M = -M^t$ . We easily check that, if  $M \neq 0$ , then the points  $x$ , solutions of  $b + Mx = 0$ , form a line in  $\mathbb{R}^3$  and a point in  $\mathbb{R}^2$ . Now  $v(x) = 0$  on  $\partial\Omega_D$ , which has nonzero surface measure, therefore  $M = 0$  and  $b = 0$ . Let us now show (5.65) by contradiction (see remark 4.3.18). If (5.65) is false, there exists a sequence  $v_n \in V$  such that

$$\|v_n\|_{H^1(\Omega)} = 1 > n \|e(v_n)\|_{L^2(\Omega)}.$$

In particular, the sequence  $e(v_n)$  tends to 0 in  $L^2(\Omega)^{N^2}$ . On the other hand, as  $v_n$  is bounded in  $H^1(\Omega)^N$ , by application of the Rellich theorem 4.3.21, there exists a subsequence  $v_{n'}$  which converges in  $L^2(\Omega)^N$ . Korn's inequality in lemma 5.3.3 implies that

$$\|v_{n'} - v_{p'}\|_{H^1(\Omega)}^2 \leq C \|v_{n'} - v_{p'}\|_{L^2(\Omega)}^2 + \|e(v_{n'}) - e(v_{p'})\|_{L^2(\Omega)}^2,$$

from which we deduce that the sequence  $v_{n'}$  is Cauchy in  $H^1(\Omega)^N$ , and therefore converges to a limit  $v_{\infty}$  which satisfies  $\|e(v_{\infty})\|_{L^2(\Omega)} = 0$ . As this is a norm we deduce that the limit is zero,  $v_{\infty} = 0$ , which is a contradiction with the fact that  $\|v_{n'}\|_{H^1(\Omega)} = 1$ .

The interpretation of the variational formulation (5.64) to recover the equation (5.60) is similar to that we have made in the proof of the theorem 5.2.14 on the Laplacian with Neumann boundary conditions. Finally, the mapping  $(f, g) \rightarrow u$  is linear. To show that it is continuous from  $L^2(\Omega)^N \times L^2(\partial\Omega)^N$  into  $H^1(\Omega)^N$ , we take  $v = u$  in the variational formulation (5.64). By using the coercivity of the bilinear form and by bounding the linear form above, we obtain **the energy estimate**

$$C\|u\|_{H^1(\Omega)}^2 \leq \|f\|_{L^2(\Omega)}\|u\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega_N)}\|u\|_{L^2(\partial\Omega)}. \quad (5.66)$$

Thanks to the Poincaré inequality and to the trace theorem, we can bound the term on the right of (5.66) above by  $C(\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega_N)})\|u\|_{H^1(\Omega)}$ , which proves the continuity.  $\square$

**Remark 5.3.5** For the elasticity system with Dirichlet (or Neumann) boundary conditions, we have the same regularity results as with the Laplacian (see theorem 5.2.26). Conversely, as opposed to the Laplacian, there is no maximum principle for the elasticity system (as for most systems of several equations). We give a numerical counterexample in Figure 5.3: in the absence of forces,  $f = 0$ , the boundary conditions are Neumann on the top and bottom faces of the domain, and Dirichlet with  $u = 0$  on the right and  $u = e_1$  on the left. This amounts to stretching the domain therefore leading to a vertical displacement which changes sign.  $\bullet$

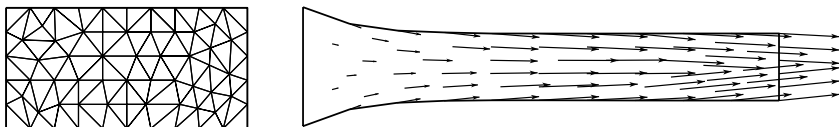


Figure 5.3. Numerical counterexample for the maximum principle in elasticity. On the left, the domain is at rest, and on the right, the domain deforms where the arrows represent the displacement (its vertical component changes sign).

We have already said that the variational formulation is nothing other than the **principle of virtual work** in mechanics. Following this analogy, the space  $V$  is the space of **kinematically admissible displacements**  $v$ , and the space of symmetric tensors  $\sigma \in L^2(\Omega)^{N^2}$ , such that  $-\operatorname{div}\sigma = f$  in  $\Omega$  and  $\sigma n = g$  on  $\partial\Omega_N$  is that of tensors of **statically admissible stress tensors**. As for the Laplacian, the solution of the variational formulation (5.64) attains the minimum of a mechanical energy defined for  $v \in V$  by

$$J(v) = \frac{1}{2} \int_{\Omega} (2\mu|e(v)|^2 + \lambda|\operatorname{div}v|^2) dx - \int_{\Omega} f \cdot v dx - \int_{\partial\Omega_N} g \cdot v ds. \quad (5.67)$$

In mechanical terms,  $J(v)$  is the sum of **the energy of deformation**

$$\frac{1}{2} \int_{\Omega} (2\mu|e(v)|^2 + \lambda|\operatorname{div}v|^2) dx$$

and of **the potential energy of exterior forces** (or work of exterior forces up to a given sign)

$$-\int_{\Omega} f \cdot v \, dx - \int_{\partial\Omega_N} g \cdot v \, ds.$$

**Exercise 5.3.3** Show that  $u \in V$  is the unique solution of the variational formulation (5.64), if and only if  $u$  attains the minimum over  $V$  of the energy  $J(v)$  defined by (5.67). (Hint: as a starting point we can take proposition 3.3.4).

**Exercise 5.3.4** Let  $\Omega$  be an open bounded connected set of  $\mathbb{R}^N$ . We consider the elasticity system with the Neumann condition (5.59) over all the boundary  $\partial\Omega$ . Show that the equilibrium condition

$$\int_{\Omega} f \cdot (Mx + b) \, dx + \int_{\partial\Omega} g \cdot (Mx + b) \, ds = 0 \quad \forall b \in \mathbb{R}^N, \quad \forall M = -M^t \in \mathbb{R}^{N \times N}$$

is a necessary and sufficient condition for existence and uniqueness of a solution in  $H^1(\Omega)^N$  (the uniqueness being obtained up to the addition of a given 'rigid body motion', see (5.62)).

**Remark 5.3.6** When the Lamé coefficients are constant and the boundary conditions are Dirichlet and homogeneous, the elasticity equations may be rearranged to give the Lamé system (presented in Chapter 1)

$$\begin{cases} -\mu\Delta u - (\mu + \lambda)\nabla(\operatorname{div} u) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (5.68)$$

The advantage of (5.68) with respect to (5.56) is that the tensor  $e(u)$  has disappeared and that we can therefore do without Korn's inequality. We should, however, draw attention to the fact that (5.68) is no longer equivalent to (5.56) if the Lamé coefficients depend on  $x$  or if we have Neumann conditions on a part of the boundary. From a mechanical point of view, the only correct model in this case is (5.56). •

**Exercise 5.3.5** We assume that  $\Omega$  is an open bounded set of  $\mathbb{R}^N$  and that  $f \in L^2(\Omega)^N$ . Show the existence and uniqueness of the solution of (5.68) in  $H_0^1(\Omega)^N$  without using Korn's inequality. Verify that we can weaken the hypotheses of positivity on the Lamé coefficients by assuming only that  $\mu > 0$  and  $2\mu + \lambda > 0$ .

**Exercise 5.3.6** Verify the equivalence of (5.68) and (5.56) if  $\lambda$  and  $\mu$  are constants. Show that (5.68) and (5.56) are no longer equivalent if  $\lambda$  and  $\mu$  are (regular) functions, even if we replace the vector equation (5.68) by

$$-\operatorname{div}(\mu\nabla u) - \nabla((\mu + \lambda)\operatorname{div} u) = f \text{ in } \Omega.$$

In a very particular case, called the **problem of antiplane shearing**, the system of linear elasticity simplifies considerably as it reduces to the solution of a boundary value problem for the Laplacian. This example therefore allows us to make a direct link between the elasticity equations and the Laplacian, which explains why the results for the two models are very similar overall. This particular case of antiplane shearing is studied in the following exercise.

**Exercise 5.3.7** The aim of this exercise is to find a particular solution of the system of linear elasticity in the case of an antiplane shearing force. We consider a homogeneous cylindrical domain  $\Omega$  of length  $L > 0$  and with section  $\omega$ , where  $\omega$  is a connected regular open bounded set of  $\mathbb{R}^{N-1}$  (the Lamé coefficients  $\lambda$  and  $\mu$  are constants). In other words,  $\Omega = \omega \times (0, L)$ , and for  $x \in \Omega$ , we denote  $x = (x', x_N)$  with  $0 < x_N < L$  and  $x' \in \omega$ . We consider the following boundary value problem

$$\begin{cases} -\operatorname{div}(2\mu e(u) + \lambda \operatorname{tr}(e(u)) \mathbf{I}) = 0 & \text{in } \Omega \\ \sigma n = g & \text{on } \partial\omega \times (0, L) \\ u' = 0 & \text{on } \omega \times \{0, L\} \\ (\sigma n) \cdot n = 0 & \text{on } \omega \times \{0, L\} \end{cases} \quad (5.69)$$

where we have used the notation, for a vector  $v = (v_1, \dots, v_N)$ ,  $v = (v', v_N)$  with  $v' \in \mathbb{R}^{N-1}$  and  $v_N \in \mathbb{R}$ . We assume that the surface force  $g$  is of the 'antiplane shearing' type, that is  $g' = (g_1, \dots, g_{N-1}) = 0$  and  $g_N$  only depends on  $x'$ .

Show that the unique solution of (5.69) is given by  $u = (0, \dots, 0, u_N)$  where  $u_N(x')$  is the solution of the following Laplacian

$$\begin{cases} -\Delta' u_N = 0 & \text{in } \omega \\ \mu \frac{\partial u_N}{\partial n} = g_N & \text{on } \partial\omega \end{cases}$$

where  $\Delta'$  is the Laplacian in the variable  $x' \in \mathbb{R}^{N-1}$ .

**Exercise 5.3.8** Generalize exercise 5.3.7 to the case of a lateral boundary condition of the type

$$u' = 0 \quad \text{and} \quad (\sigma n) \cdot e_N = g_N \quad \text{on } \partial\omega \times (0, L).$$

**Exercise 5.3.9** With the help of the variational approach show the existence and uniqueness of the solution of the plate equation

$$\begin{cases} \Delta(\Delta u) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega \end{cases} \quad (5.70)$$

where  $f \in L^2(\Omega)$ . Hint: we notice that, if  $u \in H_0^2(\Omega)$ , then  $\frac{\partial u}{\partial x_i} \in H_0^1(\Omega)$  and

$$\int_{\Omega} |\Delta u|^2 dx = \sum_{i,j=1}^N \int_{\Omega} \left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 dx.$$



We admit the following regularity result: if  $w \in L^2(\Omega)$  and  $f \in L^2(\Omega)$  satisfy for all  $v \in C_c^\infty(\Omega)$

$$-\int_{\Omega} w \Delta v \, dx = \int_{\Omega} f v \, dx,$$

then  $(\theta w) \in H^2(\Omega)$  whatever the function  $\theta \in C_c^\infty(\Omega)$ .

### 5.3.2 Stokes equations

We apply the variational approach to the solution of the system of Stokes equations. Let  $\Omega$  be a connected open bounded set of  $\mathbb{R}^N$ . Let a force  $f(x)$ , be a function from  $\Omega$  into  $\mathbb{R}^N$ . There are two unknowns: the velocity  $u$  which is a vector function, and the pressure  $p$  which is a scalar function. In vector notation, the boundary value problem considered is

$$\begin{cases} \nabla p - \mu \Delta u = f & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (5.71)$$

where  $\mu > 0$  is the viscosity of the fluid. The second equation of (5.71) is the incompressibility constraint for the fluid, while the first is the balance of forces. The Dirichlet boundary conditions model the adherence of the fluid to the walls.

**Remark 5.3.7** The Stokes problem (5.71) is a **simplified** model of the flow of a viscous incompressible fluid (in a stationary regime). In effect, the ‘real’ equations for the movement of such a fluid are the stationary Navier–Stokes equations (see, for example, [39])

$$\begin{cases} (u \cdot \nabla)u + \nabla p - \mu \Delta u = f & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (5.72)$$

When the velocity of the fluid  $u$  is small, the nonlinear term  $(u \cdot \nabla)u$ , being quadratic in  $u$ , becomes negligible. We then obtain the Stokes equations. We must therefore be aware that the domain of validity of this model is limited by this hypothesis of small velocity. •

**Theorem 5.3.8** *Let  $\Omega$  be a regular open bounded connected set of class  $\mathcal{C}^1$  of  $\mathbb{R}^N$ . Take  $f \in L^2(\Omega)^N$ . There exists a unique (weak) solution  $u \in H_0^1(\Omega)^N$  and  $p \in L^2(\Omega)/\mathbb{R}$  of (5.71) (the pressure is unique up to an additive constant in  $\Omega$ ).*

**Proof.** To find the variational formulation we multiply each equation of the system (5.71) by a test function  $v_i$  (which is zero on the boundary  $\partial\Omega$  to take account of the Dirichlet boundary conditions), we integrate by parts and we sum for  $i$  going from 1 to  $N$  (to obtain the divergence of the function  $v = (v_1, \dots, v_N)$ )

$$\int_{\Omega} \mu \nabla u \cdot \nabla v \, dx - \int_{\Omega} p \operatorname{div} v \, dx = \int_{\Omega} f \cdot v \, dx,$$

with the notation  $\nabla u \cdot \nabla v = \sum_{i=1}^N \nabla u_i \cdot \nabla v_i$ . Taking into account the incompressibility condition  $\operatorname{div} u = 0$ , we choose as the Hilbert space the following subspace of  $H_0^1(\Omega)^N$

$$V = \{v \in H_0^1(\Omega)^N \text{ such that } \operatorname{div} v = 0 \text{ a.e. in } \Omega\}, \quad (5.73)$$

which is a Hilbert space as a closed subspace of  $H_0^1(\Omega)^N$ . We then find the variational formulation:

$$\text{find } u \in V \text{ such that } \int_{\Omega} \mu \nabla u \cdot \nabla v \, dx = \int_{\Omega} f \cdot v \, dx \quad \forall v \in V, \quad (5.74)$$

in which the pressure has disappeared! We see easily that each term of (5.74) has a meaning.

The application of the Lax–Milgram theorem to the variational formulation (5.74) poses no problem. In particular, the coercivity of the bilinear form is obvious in  $H_0^1(\Omega)^N$  (thanks to the Poincaré inequality) therefore in  $V$  which is a subspace of  $H_0^1(\Omega)^N$ .

The most delicate point here is to show that the unique solution of (5.74) is a solution of the boundary value problem (5.71). Let us explain the difficulty while assuming temporarily that  $u$  is regular, that is belongs to  $H^2(\Omega)^N$ . By integration by parts, (5.74) implies that

$$\int_{\Omega} (\mu \Delta u + f) \cdot v \, dx = 0 \quad \forall v \in V,$$

but we cannot deduce that  $(\mu \Delta u + f) = 0$  since the orthogonal complement of  $V$  in  $L^2(\Omega)^N$  is not empty! In effect, we see easily that if  $\phi$  is a regular function, then, for all  $v \in V$ , we have

$$\int_{\Omega} \nabla \phi \cdot v \, dx = - \int_{\Omega} \phi \operatorname{div} v \, dx = 0,$$

that is the orthogonal complement of  $V$  contains at least all the gradients. In fact de Rham's theorem 5.3.9 tells us that the orthogonal complement of  $V$  coincides exactly with the space of gradients.

Thanks to de Rham's theorem 5.3.9 we can finish as follows. We set

$$L(v) = \int_{\Omega} \mu \nabla u \cdot \nabla v \, dx - \int_{\Omega} f \cdot v \, dx,$$

which is a continuous linear form over  $H_0^1(\Omega)^N$  and zero over  $V$ . Consequently, there exists  $p \in L^2(\Omega)$ , unique up to a given constant, such that

$$L(v) = \int_{\Omega} p \operatorname{div} v \, dx \quad \forall v \in H_0^1(\Omega)^N.$$

If we set  $\sigma = \mu \nabla u - p \mathbf{I}$  which belongs to  $L^2(\Omega)^{N^2}$ , we therefore have

$$\left| \int_{\Omega} \sigma \cdot \nabla v \, dx \right| = \left| \int_{\Omega} f \cdot v \, dx \right| \leq C \|v\|_{L^2(\Omega)},$$

which proves that  $\sigma$  has a weak divergence in  $L^2(\Omega)^N$ , and we have  $-\operatorname{div} \sigma = f$ . Consequently, we deduce that

$$\nabla p - \mu \Delta u = f \text{ almost everywhere in } \Omega. \quad (5.75)$$

On the other hand, as  $u \in V$ ,  $\operatorname{div} u$  is zero in  $L^2(\Omega)$ , this implies

$$\operatorname{div} u = 0 \text{ almost everywhere in } \Omega.$$

Likewise, the boundary conditions are interpreted by the trace theorem and we obtain  $u = 0$  almost everywhere on  $\partial\Omega$ .  $\square$

We now state the very profound and difficult result which allows us to recover the pressure from the variational formulation (5.74) of the Stokes system where it has disappeared! Its proof is far beyond the framework of this course (note in passing that it is not easy to find in the literature an ‘elementary’ and self-contained proof; see nevertheless [21]).

**Theorem 5.3.9 (Rham)** *Let  $\Omega$  be a regular open bounded connected set of class  $\mathcal{C}^1$  of  $\mathbb{R}^N$ . Let  $L$  be a continuous linear form over  $H_0^1(\Omega)^N$ . Then  $L$  is zero on  $V$  if and only if there exists a function  $p \in L^2(\Omega)$  such that*

$$L(v) = \int_{\Omega} p \operatorname{div} v \, dx \quad \forall v \in H_0^1(\Omega)^N. \quad (5.76)$$

*Further  $p$  is unique up to a given additive constant.*

**Remark 5.3.10** For Stokes as for the Laplacian or for elasticity we can also define Neumann boundary conditions which are written

$$\mu \frac{\partial u}{\partial n} - pn = g \quad \text{on } \partial\Omega, \quad (5.77)$$

where  $g$  is a (vector valued) function of  $L^2(\partial\Omega)^N$ . As for elasticity, the Neumann condition (5.77) is interpreted by saying that  $g$  is a force applied on the boundary. •

**Remark 5.3.11** For the Stokes system (5.71) we have the same regularity results as for the Laplacian or the system of linear elasticity (see theorem 5.2.26). Conversely, the Stokes equations being a system of several equations, there is no maximum principle (this is the same situation as for the elasticity system). Physically, because of the incompressibility condition of the fluid we understand this very well. In effect, if we study a Stokes flow in a duct with a pronounced narrowing, as the flux is constant across a section, the velocity of the fluid is necessarily higher in the narrow part than at the entrance or exit (which violates the maximum principle in the absence of exterior forces). •

As for the elasticity system there exists a principle of minimization of the energy (called **viscous dissipation**) for the Stokes equations.

**Exercise 5.3.10** Let  $V$  be the space velocity field with zero divergence defined by (5.73). Let  $J(v)$  be the energy defined for  $v \in V$  by

$$J(v) = \frac{1}{2} \int_{\Omega} \mu |\nabla v|^2 dx - \int_{\Omega} f \cdot v dx. \quad (5.78)$$

Let  $u \in V$  be the unique solution of the variational formulation (5.74). Show that  $u$  is also the unique point of minimum of the energy, that is,  $J(u) = \min_{v \in V} J(v)$ . Conversely, show that, if  $u \in V$  is a minimum point of the energy  $J(v)$ , then  $u$  is the unique solution of the variational formulation (5.74).

In the following exercise we shall see that in certain very particular cases the Stokes equations reduce to the Laplacian.

**Exercise 5.3.11** The aim of this exercise is to find a particular solution of the Stokes equations in a rectangular channel with uniform section, called the Poiseuille profile. Take  $\Omega = \omega \times (0, L)$  where  $L > 0$  is the length of the channel and  $\omega$  its section, a regular open bounded connected set of  $\mathbb{R}^{N-1}$ . For  $x \in \Omega$ , we denote  $x = (x', x_N)$  with  $0 < x_N < L$  and  $x' \in \omega$ . We consider the following boundary value problem

$$\begin{cases} \nabla p - \mu \Delta u = 0 & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\omega \times (0, L) \\ pn - \mu \frac{\partial u}{\partial n} = p_0 n & \text{on } \omega \times \{0\} \\ pn - \mu \frac{\partial u}{\partial n} = p_L n & \text{on } \omega \times \{L\} \end{cases} \quad (5.79)$$

where  $p_0$  and  $p_L$  are two constant pressures. Show that the unique solution of (5.79) is  $p(x) = p_0 + \frac{x_N}{L}(p_L - p_0)$ , and  $u = (0, \dots, 0, u_N)$  where  $u_N$  is the solution of the following Laplacian

$$\begin{cases} -\mu \Delta' u_N = -(p_L - p_0)/L & \text{in } \omega \\ u_N = 0 & \text{on } \partial\omega \end{cases}$$

where  $\Delta'$  is the Laplacian in the variable  $x' \in \mathbb{R}^{N-1}$ .

**Exercise 5.3.12** Generalize exercise 5.3.11 to the case of Navier–Stokes equations (5.72).

*This page intentionally left blank*

# 6 Finite element method

---

## 6.1 Variational approximation

### 6.1.1 Introduction

In this chapter, we present the method of **finite elements** which is the numerical method of choice for the calculation of solutions of elliptic boundary value problems, but is also used for parabolic or hyperbolic problems as we shall see. The principle of this method comes directly from the **variational approach** that we have studied in detail in the preceding chapters.

The idea at the base of the finite element method is to replace the Hilbert space  $V$  on which we pose the variational formulation by a subspace  $V_h$  of finite dimension. The ‘approximate’ problem posed over  $V_h$  reduces to the simple solution of a linear system, whose matrix is called the **stiffness matrix**. In addition, we can choose the construction of  $V_h$  in such a way that the subspace  $V_h$  is a good approximation of  $V$  and that the solution  $u_h$  in  $V_h$  of the variational formulation is ‘**close**’ to the exact solution  $u$  in  $V$ .

Historically, the first premises of the finite element method have been proposed by the mathematician Richard Courant (without using this name) in the 1940s, but it was mechanical engineers who have developed, popularized, and proved the efficiency of this method in the 1950s and 1960s (as well as giving it its actual name). After these first practical successes, mathematicians have then considerably developed the theoretical foundations of the method and proposed significant improvements. This is in any case a good example of interdisciplinary cooperation where the joint efforts of engineers and applied mathematicians have made immense progress in numerical simulation (not neglecting the even more spectacular advances in the power of computers).

The plan of this chapter is the following. In the rest of this section we detail the process of **internal variational approximation**. Section 6.2 presents finite elements in one space dimension where, without betraying the general ideas valid in higher dimensions, the technical aspects are much simpler. We discuss some practical

aspects (assembly of the stiffness matrix, quadrature formulas, etc.) as much as the theoretical (convergence of the method, interpolation and error estimation). Section 6.3 is dedicated to finite elements in higher dimensions ( $N \geq 2$ ). We introduce the concepts of the **mesh** (triangular or quadrilateral) and of **degrees of freedom** which will allow us to construct several families of finite element methods. We then look at the practical and theoretical aspects already discussed in  $N = 1$  dimension.

Let us finish this introduction by saying that as well as the finite difference and finite element methods, which are the only ones discussed in this course, there exist other numerical methods for the solution of partial differential equations like the finite volume method, the boundary element method, the spectral method, the Fourier method, etc. (see the encyclopaedias [10], [14]). For more details on the finite element method we refer to [5], [9], [17], [21], [32], [34] (see also [33], [13], [29] for practical aspects of computer programming).

### 6.1.2 General internal approximation

We again consider the general framework of the variational formalism introduced in Chapter 3. Given a Hilbert space  $V$ , a continuous and coercive bilinear form  $a(u, v)$ , and a continuous linear form  $L(v)$ , we consider the variational formulation:

$$\text{find } u \in V \text{ such that } a(u, v) = L(v) \quad \forall v \in V, \quad (6.1)$$

which we know has a unique solution by the Lax–Milgram theorem 3.3.1. The **internal approximation** of (6.1) consists of replacing the Hilbert space  $V$  by a finite dimensional subspace  $V_h$ , that is to look for the solution of:

$$\text{find } u_h \in V_h \text{ such that } a(u_h, v_h) = L(v_h) \quad \forall v_h \in V_h. \quad (6.2)$$

The solution of the internal approximation (6.2) is easy as we show in the following lemma.

**Lemma 6.1.1** *Let  $V$  be a real Hilbert space, and  $V_h$  a finite dimensional subspace. Let  $a(u, v)$  be a continuous and coercive bilinear form over  $V$ , and  $L(v)$  a continuous linear form over  $V$ . Then the internal approximation (6.2) has a unique solution. In addition, this solution can be obtained by solving a linear system with a positive definite matrix (and symmetric if  $a(u, v)$  is symmetric).*

**Proof.** The existence and uniqueness of  $u_h \in V_h$ , the solution of (6.2), follows from the Lax–Milgram theorem 3.3.1 applied to  $V_h$ . To put the problem in a simpler form, we introduce a basis  $(\phi_j)_{1 \leq j \leq N_h}$  of  $V_h$ . If  $u_h = \sum_{j=1}^{N_h} u_j \phi_j$ , we set  $U_h = (u_1, \dots, u_{N_h})$  the vector in  $\mathbb{R}^{N_h}$  of coordinates of  $u_h$ . Problem (6.2) is equivalent to:

$$\text{find } U_h \in \mathbb{R}^{N_h} \text{ such that } a \left( \sum_{j=1}^{N_h} u_j \phi_j, \phi_i \right) = L(\phi_i) \quad \forall 1 \leq i \leq N_h,$$

which can be written in the form of a linear system

$$\mathcal{K}_h U_h = b_h, \quad (6.3)$$

with, for  $1 \leq i, j \leq N_h$ ,

$$(\mathcal{K}_h)_{ij} = a(\phi_j, \phi_i), \quad (b_h)_i = L(\phi_i).$$

The coercivity of the bilinear form  $a(u, v)$  implies the positive definite character of the matrix  $\mathcal{K}_h$ , and therefore its invertibility. In effect, for every vector  $U_h \in \mathbb{R}^{N_h}$ , we have

$$\mathcal{K}_h U_h \cdot U_h \geq \nu \left\| \sum_{j=1}^{N_h} u_j \phi_j \right\|^2 \geq C |U_h|^2 \quad \text{with } C > 0,$$

since all norms are equivalent in finite dimensions ( $|\cdot|$  denotes the Euclidean norm in  $\mathbb{R}^{N_h}$ ). Likewise, the symmetry of  $a(u, v)$  implies that of  $\mathcal{K}_h$ . In engineering applications the matrix  $\mathcal{K}_h$  is called the **stiffness matrix**.  $\square$

We shall now compare the error caused by replacing the space  $V$  by its subspace  $V_h$ . More precisely, we shall bound the difference  $\|u - u_h\|$  where  $u$  is the solution in  $V$  of (6.1) and  $u_h$  that in  $V_h$  of (6.2). Let us first make some notation precise: we denote by  $\nu > 0$  the coercivity constant and  $M > 0$  the continuity constant of the bilinear form  $a(u, v)$  which satisfy

$$\begin{aligned} a(u, u) &\geq \nu \|u\|^2 \quad \forall u \in V, \\ |a(u, v)| &\leq M \|u\| \|v\| \quad \forall u, v \in V. \end{aligned}$$

The following lemma, due to Jean Céa, shows that the distance between the exact solution  $u$  and the approximate solution  $u_h$  is bounded **uniformly with respect to the subspace**  $V_h$  by the distance between  $u$  and  $V_h$ .

**Lemma 6.1.2 (Céa)** *We use the hypotheses of lemma 6.1.1. Let  $u$  be the solution of (6.1) and  $u_h$  that of (6.2). We have*

$$\|u - u_h\| \leq \frac{M}{\nu} \inf_{v_h \in V_h} \|u - v_h\|. \quad (6.4)$$

**Proof.** Because  $V_h \subset V$ , we deduce, by subtraction of the variational formulations (6.1) and (6.2), that

$$a(u - u_h, w_h) = 0 \quad \forall w_h \in V_h.$$

By choosing  $w_h = u_h - v_h$  we obtain

$$\nu \|u - u_h\|^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \leq M \|u - u_h\| \|u - v_h\|,$$

from which we deduce (6.4).  $\square$



**Exercise 6.1.1** In the framework of Céa's lemma 6.1.2, prove that, if the bilinear form  $a(u, v)$  is symmetric, then we improve (6.4) to

$$\|u - u_h\| \leq \sqrt{\frac{M}{\nu}} \inf_{v_h \in V_h} \|u - v_h\|.$$

Hint: use the fact that the solution  $u_h$  of (6.2) also attains the minimum of an energy.

Finally, to prove the convergence of this variational approximation, we give a last general lemma. Recall that in the notation  $V_h$  the parameter  $h > 0$  does not have a practical meaning. Nevertheless, we shall assume that it is in the limit  $h \rightarrow 0$  that the internal approximation (6.2) 'converges' to the variational formulation (6.1).

**Lemma 6.1.3** *We use the hypotheses of lemma 6.1.1. We assume that there exists a subspace  $\mathcal{V} \subset V$  which is dense in  $V$  and a mapping  $r_h$  from  $\mathcal{V}$  into  $V_h$  (called an **interpolation operator**) such that*

$$\lim_{h \rightarrow 0} \|v - r_h(v)\| = 0 \quad \forall v \in \mathcal{V}. \quad (6.5)$$

*Then the method of internal variational approximation converges, that is,*

$$\lim_{h \rightarrow 0} \|u - u_h\| = 0. \quad (6.6)$$

**Proof.** Take  $\epsilon > 0$ . By density of  $\mathcal{V}$ , there exists  $v \in \mathcal{V}$  such that  $\|u - v\| \leq \epsilon$ . In addition, there exists an  $h_0 > 0$  (depending on  $\epsilon$ ) such that, for this element  $v \in \mathcal{V}$ , we have

$$\|v - r_h(v)\| \leq \epsilon \quad \forall h \leq h_0.$$

From lemma 6.1.2, we have

$$\|u - u_h\| \leq C\|u - r_h(v)\| \leq C(\|u - v\| + \|v - r_h(v)\|) \leq 2C\epsilon,$$

from which we deduce the result.  $\square$

The strategy indicated by lemmas 6.1.1, 6.1.2, and 6.1.3 above is now clear. **To obtain a numerical approximation of the exact solution of the variational problem (6.1), we must introduce a finite dimensional space  $V_h$  then solve a simple linear system associated with the internal variational approximation (6.2).** Nevertheless, the choice of  $V_h$  is not obvious. It must satisfy two criteria:

1. We must construct an interpolation operator  $r_h$  from  $\mathcal{V}$  into  $V_h$  satisfying (6.5) (where typically  $\mathcal{V}$  is a space of regular functions).
2. The solution of the linear system  $\mathcal{K}_h U_h = b_h$  must be economical (in practice these linear systems are very large).

The finite element method consists precisely of providing such 'good' spaces as  $V_h$ . Before entering into the details, we shall say something about the Galerkin method which appears in this framework.

### 6.1.3 Galerkin method

The Galerkin method has been a precursor of the finite element method. Even though it does not have any numerical interest in general, it is very useful from a theoretical point of view (notably for the study of nonlinear problems). It appears in the framework of the internal variational approximation described above.

We assume that the Hilbert space  $V$  is separable and infinite dimensional, which implies, by proposition 12.1.15, that there exists a Hilbertian basis  $(e_i)_{i \geq 1}$  of  $V$ . We then choose  $\mathcal{V}$  as the subspace generated by this Hilbertian basis (generated by a finite linear combination) which is dense in  $V$ . By setting  $h = 1/n$ , we define  $V_h$  as the finite dimensional subspace generated by  $(e_1, \dots, e_n)$ . Finally, the interpolation operator  $r_h$  is simply the orthogonal projection over  $V_h$  (which is here defined in all of  $V$  and not only in  $\mathcal{V}$ ).

All the hypotheses of lemmas 6.1.1, 6.1.2, and 6.1.3 are therefore satisfied and we deduce that the approximate solution  $u_h$  converges to the exact solution  $u$ . Recall that  $u_h$  is calculated by solving the linear system  $\mathcal{K}_h U_h = b_h$  where  $U_h$  is the vector in  $\mathbb{R}^n$  of coordinates of  $u_h$  in the basis  $(e_1, \dots, e_n)$ .

Despite its usefulness in the theoretical framework developed above, the Galerkin method is not helpful from a numerical point of view. In effect, the matrix  $\mathcal{K}_h$  that we obtain is generally ‘full’, that is, all the coefficients are nonzero in general, and ‘ill-conditioned’, that is, the numerical solution of the linear system will be unstable and so very sensitive to rounding errors on the computer. From this point of view, the finite element method is much more powerful and far preferable to the Galerkin method.

### 6.1.4 Finite element method (general principles)

The principle of the finite element method is to construct internal approximation spaces  $V_h$  from the usual functional spaces  $H^1(\Omega), H_0^1(\Omega), H^2(\Omega), \dots$ , whose definition is based on the geometrical concept of a **mesh** of the domain  $\Omega$ . A mesh is a tessellation of the space by very simple elementary volumes: triangles, tetrahedra, parallelopipeds (see, for example, Figure 6.7). We shall later give a precise definition of a mesh in the framework of the finite element method.

In this context the parameter  $h$  of  $V_h$  corresponds to the **maximum size of the mesh** or the cells which comprise the mesh. Typically a basis of  $V_h$  will be composed of functions whose support is **localized** in one or few elements. This will have two important consequences: on the one hand, in the limit  $h \rightarrow 0$ , the space  $V_h$  will be more and more ‘large’ and will approach little by little the entire space  $V$ , and on the other hand, the stiffness matrix  $\mathcal{K}_h$  of the linear system (6.3) will be **sparse**, that is, most of its coefficients will be zero (which will limit the cost of the numerical solution).

The finite element method is one of the most effective and most popular methods of numerically solving boundary value problems. It is the basis of innumerable industrial software packages.

## 6.2 Finite elements in $N = 1$ dimension

To simplify the exposition, we start by presenting the finite element method in one space dimension. Without loss of generality we choose the domain  $\Omega = ]0, 1[$ . In one dimension a mesh is simply composed of a collection of points  $(x_j)_{0 \leq j \leq n+1}$  (as for the finite difference method, see Chapter 1) such that

$$x_0 = 0 < x_1 < \cdots < x_n < x_{n+1} = 1.$$

The mesh will be called **uniform** if the points  $x_j$  are equidistant, that is,

$$x_j = jh \quad \text{with } h = \frac{1}{n+1}, \quad 0 \leq j \leq n+1.$$

The points  $x_j$  are also called the **vertices** or nodes of the mesh. For simplicity we consider, for the moment, the following model problem

$$\begin{cases} -u'' = f & \text{in } ]0, 1[ \\ u(0) = u(1) = 0, \end{cases} \quad (6.7)$$

which we know has a unique solution in  $H_0^1(\Omega)$  if  $f \in L^2(\Omega)$  (see Chapter 5). In all that follows, we denote by  $\mathbb{P}_k$  the set of polynomials, with real coefficients, of one real variable with degree less than or equal to  $k$ .

### 6.2.1 $\mathbb{P}_1$ finite elements

The  $\mathbb{P}_1$  finite element method uses the discrete space of globally continuous functions which are affine on each element

$$V_h = \{v \in C([0, 1]) \quad \text{such that } v|_{[x_j, x_{j+1}]} \in \mathbb{P}_1 \text{ for all } 0 \leq j \leq n\}, \quad (6.8)$$

and on its subspace

$$V_{0h} = \{v \in V_h \quad \text{such that } v(0) = v(1) = 0\}. \quad (6.9)$$

The  $\mathbb{P}_1$  finite element method is then simply the method of internal variational approximation of Section 6.1.2 applied to the spaces  $V_h$  or  $V_{0h}$  defined by (6.8) or (6.9).

We can represent the functions of  $V_h$  or  $V_{0h}$ , which are piecewise affine, with the help of very simple basis functions. We introduce the ‘hat function’  $\phi$  defined by

$$\phi(x) = \begin{cases} 1 - |x| & \text{if } |x| \leq 1, \\ 0 & \text{if } |x| > 1. \end{cases}$$

If the mesh is uniform, for  $0 \leq j \leq n+1$  we define the basis functions (see Figure 6.1)

$$\phi_j(x) = \phi\left(\frac{x - x_j}{h}\right). \quad (6.10)$$

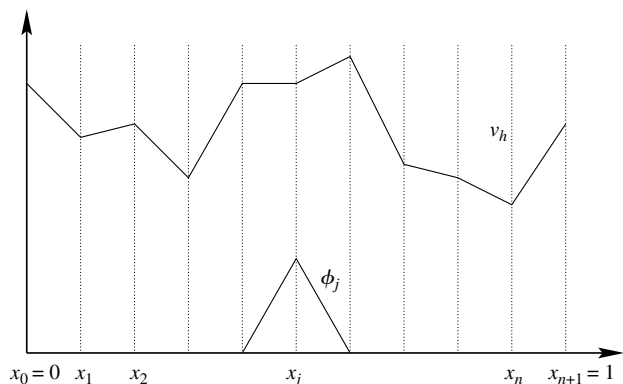


Figure 6.1. Mesh of  $\Omega = ]0, 1[$  and  $\mathbb{P}_1$  finite element basis functions.

**Lemma 6.2.1** *The space  $V_h$ , defined by (6.8), is a subspace of  $H^1(0, 1)$  of dimension  $n + 2$ , and every function  $v_h \in V_h$  is defined uniquely by its values at the vertices  $(x_j)_{0 \leq j \leq n+1}$*

$$v_h(x) = \sum_{j=0}^{n+1} v_h(x_j) \phi_j(x) \quad \forall x \in [0, 1].$$

*Likewise,  $V_{0h}$ , defined by (6.9), is a subspace of  $H_0^1(0, 1)$  of dimension  $n$ , and every function  $v_h \in V_{0h}$  is defined uniquely by its values at the vertices  $(x_j)_{1 \leq j \leq n}$*

$$v_h(x) = \sum_{j=1}^n v_h(x_j) \phi_j(x) \quad \forall x \in [0, 1].$$

**Proof.** Let us recall that by virtue of lemma 4.3.19, the continuous functions which are piecewise of class  $C^1$  belong to  $H^1(\Omega)$ . Therefore,  $V_h$  and  $V_{0h}$  are subspaces of  $H^1(0, 1)$ . The rest of the proof is immediate by remarking that  $\phi_j(x_i) = \delta_{ij}$ , where  $\delta_{ij}$  is the Kronecker delta which is 1 if  $i = j$  and 0 otherwise (see Figure 6.1).  $\square$

**Remark 6.2.2** The basis  $(\phi_j)$ , defined by (6.10), allows us to characterize a function of  $V_h$  by its values at the nodes of the mesh. In this case we talk of **Lagrange finite elements**. We see later in Section 6.2.5 that we can introduce other spaces  $V_h$  for which a function will be characterized, not only by its values, but also by the values of its derivative. We talk then of **Hermite finite elements**. Here, as the functions are locally  $\mathbb{P}_1$ , we say that the space  $V_h$ , defined by (6.8), is the space of Lagrange finite elements of order 1.

This example of  $\mathbb{P}_1$  finite elements again makes it possible to understand the interest in the variational formulation. In effect, the functions of  $V_h$  are not twice

differentiable on the segment  $[0, 1]$  and this is not enough to solve, even approximately, the equation (6.7) (in fact, the second derivative of a function of  $V_h$  is a sum of Dirac masses at the nodes of the mesh!). On the contrary, it is perfectly legitimate to use functions of  $V_h$  in the variational formulation (6.2) which only requires a single derivative. •

Let us now describe the **practical solution** of the Dirichlet problem (6.7) by the  $\mathbb{P}_1$  finite element method. The variational formulation (6.2) of the internal approximation here becomes:

$$\text{find } u_h \in V_{0h} \text{ such that } \int_0^1 u'_h(x) v'_h(x) dx = \int_0^1 f(x) v_h(x) dx \quad \forall v_h \in V_{0h}. \quad (6.11)$$

We decompose  $u_h$  in the basis of  $(\phi_j)_{1 \leq j \leq n}$  and we take  $v_h = \phi_i$  which gives

$$\sum_{j=1}^n u_h(x_j) \int_0^1 \phi'_j(x) \phi'_i(x) dx = \int_0^1 f(x) \phi_i(x) dx.$$

By denoting  $U_h = (u_h(x_j))_{1 \leq j \leq n}$ ,  $b_h = \left( \int_0^1 f(x) \phi_i(x) dx \right)_{1 \leq i \leq n}$ , and by introducing the **stiffness matrix**

$$\mathcal{K}_h = \left( \int_0^1 \phi'_j(x) \phi'_i(x) dx \right)_{1 \leq i, j \leq n},$$

the variational formulation in  $V_{0h}$  reduces to solving in  $\mathbb{R}^n$  the linear system

$$\mathcal{K}_h U_h = b_h.$$

As the basis functions  $\phi_j$  have a ‘small’ support, the intersection of supports of  $\phi_j$  and  $\phi_i$  is often empty and most of the coefficients of  $\mathcal{K}_h$  are zero. A simple calculation shows that

$$\int_0^1 \phi'_j(x) \phi'_i(x) dx = \begin{cases} -h^{-1} & \text{if } j = i - 1 \\ 2h^{-1} & \text{if } j = i \\ -h^{-1} & \text{if } j = i + 1 \\ 0 & \text{otherwise} \end{cases}$$

and the matrix  $\mathcal{K}_h$  is tridiagonal

$$\mathcal{K}_h = h^{-1} \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix}. \quad (6.12)$$

To obtain the right-hand side  $b_h$  we must calculate the integrals

$$(b_h)_i = \int_{x_{i-1}}^{x_{i+1}} f(x) \phi_i(x) dx \quad \text{if } 1 \leq i \leq n.$$

The exact evaluation of the right-hand side  $b_h$  can be difficult or impossible if the function  $f$  is complicated. In practice, we use **quadrature formulas** (or numerical integration formulas) which give an approximation of the integrals  $b_h$ . For example, we can use the ‘midpoint’ formula

$$\frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} \psi(x) dx \approx \psi\left(\frac{x_{i+1} + x_i}{2}\right),$$

or the ‘trapezium’ formula

$$\frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} \psi(x) dx \approx \frac{1}{2} (\psi(x_{i+1}) + \psi(x_i)),$$

or even

$$\frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} \psi(x) dx \approx \left( \frac{1}{6} \psi(x_{i+1}) + \frac{1}{6} \psi(x_i) + \frac{2}{3} \psi\left(\frac{x_{i+1} + x_i}{2}\right) \right).$$

The first two formulas are exact for the affine functions  $\psi$ , and the third is exact for second degree polynomials (which give an exact calculation of  $b_h$  if  $f \in V_h$ ). If the function  $\psi$  is arbitrary but regular, then these formulas are simply approximations with a remainder of the order of  $\mathcal{O}(h^2)$ , of  $\mathcal{O}(h^2)$ , and of  $\mathcal{O}(h^3)$  respectively.

The solution of the linear system  $\mathcal{K}_h U_h = b_h$  is the most costly part of the method in terms of the calculation time. This is why we present in an appendix in Section 13.1 some powerful methods for the solution. Recall that the matrix  $\mathcal{K}_h$  is necessarily invertible by application of lemma 6.1.1.

**Remark 6.2.3** The stiffness matrix  $\mathcal{K}_h$  is very similar to matrices already met in the study of finite difference methods. In fact,  $h\mathcal{K}_h$  is the limit of the matrix (2.14) (multiplied by  $1/c$ ) of the implicit scheme for the solution of the heat equation when the time step tends to infinity. We see in exercise 6.2.3 that this is not a coincidence. •

**Neumann problem** The implementation of the  $\mathbb{P}_1$  finite element method for the following Neumann problem is very similar

$$\begin{cases} -u'' + au = f & \text{in } ]0, 1[ \\ u'(0) = \alpha, u'(1) = \beta. \end{cases} \quad (6.13)$$

Recall that (6.13) has a unique solution in  $H^1(\Omega)$  if  $f \in L^2(\Omega)$ ,  $\alpha, \beta \in \mathbb{R}$ , and  $a \in L^\infty(\Omega)$  such that  $a(x) \geq a_0 > 0$  a.e. in  $\Omega$  (see Chapter 5). The variational formulation (6.2) of the internal approximation here becomes: find  $u_h \in V_h$  such that

$$\int_0^1 (u'_h(x)v'_h(x) + a(x)u_h(x)v_h(x)) dx = \int_0^1 f(x)v_h(x) dx - \alpha v_h(0) + \beta v_h(1),$$

for all  $v_h \in V_h$ . By decomposing  $u_h$  in the basis of  $(\phi_j)_{0 \leq j \leq n+1}$ , the variational formulation in  $V_h$  reduces to solving in  $\mathbb{R}^{n+2}$  the linear system

$$\mathcal{K}_h U_h = b_h,$$

with  $U_h = (u_h(x_j))_{0 \leq j \leq n+1}$ , and a new stiffness matrix

$$\mathcal{K}_h = \left( \int_0^1 (\phi'_j(x) \phi'_i(x) + a(x) \phi_j(x) \phi_i(x)) dx \right)_{0 \leq i, j \leq n+1},$$

and

$$\begin{aligned} (b_h)_i &= \int_0^1 f(x) \phi_i(x) dx \quad \text{if } 1 \leq i \leq n, \\ (b_h)_0 &= \int_0^1 f(x) \phi_0(x) dx - \alpha, \\ (b_h)_{n+1} &= \int_0^1 f(x) \phi_{n+1}(x) dx + \beta. \end{aligned}$$

When  $a(x)$  is not a constant function, it is also necessary in practice to use quadrature formulae to evaluate the coefficients of the matrix  $\mathcal{K}_h$  (as we have done in the preceding example for the right-hand side  $b_h$ ).

**Exercise 6.2.1** Apply the  $\mathbb{P}_1$  finite element method to the problem

$$\begin{cases} -u'' = f & \text{in } ]0, 1[ \\ u(0) = \alpha, u(1) = \beta, \end{cases}$$

Verify that the nonhomogeneous Dirichlet boundary conditions appear in the right-hand side of the linear system which results from this.

**Exercise 6.2.2** We take again the Neumann problem (6.13) assuming that the function  $a(x) = 0$  in  $\Omega$ . Show that the matrix of the linear system of the  $\mathbb{P}_1$  finite element method is singular. Show that we can, nevertheless, solve the linear system if the data satisfy the compatibility condition

$$\int_0^1 f(x) dx = \alpha - \beta.$$

Compare this result with theorem 5.2.18.

**Exercise 6.2.3** Apply the finite difference method (see Chapter 2) to the Dirichlet problem (6.7). Verify that with a second order centred scheme, we obtain a linear system to solve with the same matrix  $\mathcal{K}_h$  (up to a multiplicative coefficient) but with a different right-hand side  $b_h$ . Repeat the question for the Neumann problem (6.13).

**Exercise 6.2.4** We consider  $(n+2)$  (aligned) point masses situated at the points  $x_j = j/(n+1)$  for  $0 \leq j \leq n+1$  and linked by springs of the same stiffness  $k > 0$ . To each point mass, we apply a longitudinal force  $f_j$ . In the hypothesis of small (longitudinal) displacements, write the total energy of the system that we must minimize (we shall discuss the case of free or fixed extremities). Interpret the search for the position of equilibrium of the system in terms of finite elements.

### 6.2.2 Convergence and error estimation

To prove the convergence of the  $\mathbb{P}_1$  finite element method in one space dimension we follow the steps outlined in Section 6.1.2. We first of all define an **interpolation operator**  $r_h$  (as in lemma 6.1.3).

**Definition 6.2.4** *The  $\mathbb{P}_1$  interpolation operator is the linear mapping  $r_h$  from  $H^1(0, 1)$  into  $V_h$  defined, for all  $v \in H^1(0, 1)$ , by*

$$(r_h v)(x) = \sum_{j=0}^{n+1} v(x_j) \phi_j(x).$$

This definition has a meaning since, by virtue of lemma 4.3.3, the functions of  $H^1(0, 1)$  are continuous and their point values are therefore well defined. The interpolant  $r_h v$  of a function  $v$  is simply the function which is piecewise affine and coincides with  $v$  on the vertices of the mesh  $x_j$  (see Figure 6.2). Let us remark that in one space dimension the interpolant is defined for every function of  $H^1(0, 1)$ , and not only for regular functions of  $H^1(0, 1)$  (which will be the case in higher dimensions).

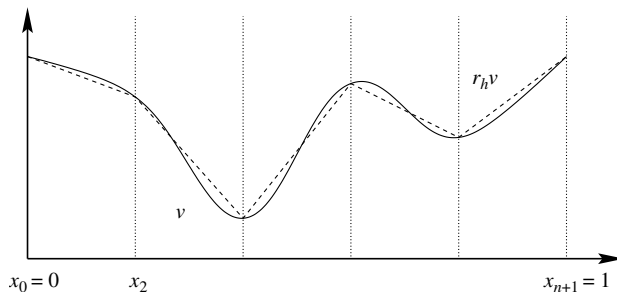


Figure 6.2.  $\mathbb{P}_1$  interpolation of a function of  $H^1(0, 1)$ .

The convergence of the  $\mathbb{P}_1$  finite element method relies on the following lemma.

**Lemma 6.2.5 (interpolation)** *Let  $r_h$  be the  $\mathbb{P}_1$  interpolation operator. For every  $v \in H^1(0, 1)$ , it satisfies*

$$\lim_{h \rightarrow 0} \|v - r_h v\|_{H^1(0,1)} = 0.$$

Moreover, if  $v \in H^2(0, 1)$ , then there exists a constant  $C$  independent of  $h$  such that

$$\|v - r_h v\|_{H^1(0,1)} \leq Ch \|v''\|_{L^2(0,1)}.$$

We temporarily delay the proof of this lemma to immediately state the principal result of this section which establishes the convergence of the finite element method  $\mathbb{P}_1$  for the Dirichlet problem.



**Theorem 6.2.6** *Let  $u \in H_0^1(0,1)$  and  $u_h \in V_{0h}$  be the solutions of (6.7) and (6.11), respectively. Then, the finite element method  $\mathbb{P}_1$  converges, that is,*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(0,1)} = 0. \quad (6.14)$$

*Moreover, if  $u \in H^2(0,1)$  (which is true if  $f \in L^2(0,1)$ ), then there exists a constant  $C$  independent of  $h$  such that*

$$\|u - u_h\|_{H^1(0,1)} \leq Ch\|u''\|_{L^2(0,1)} = Ch\|f\|_{L^2(0,1)}. \quad (6.15)$$

**Remark 6.2.7** The first conclusion (6.14) of theorem 6.2.6 is also true if the right-hand side  $f$  belongs only to  $H^{-1}(0,1)$ . The estimate (6.15) indicates the rate of convergence of the  $\mathbb{P}_1$  finite element method. As this upper bound is proportional to  $h$ , we say that the  $\mathbb{P}_1$  finite element method converges linearly.

Let us remark that the convergence theorem 6.2.6 is valid when the stiffness matrix  $\mathcal{K}_h$  and the right-hand side  $b_h$  are evaluated exactly. However, the  $\mathbb{P}_1$  finite element method also converges when we use adequate quadrature formulas to calculate  $\mathcal{K}_h$  and  $b_h$  (see [34]). •

**Remark 6.2.8** We prove theorem 6.2.6 in the case of a uniform mesh, that is, with points  $x_j$  equidistant in the segment  $[0,1]$  (in other words,  $x_{j+1} - x_j = h$ ). The result can, nevertheless, be generalized to nonuniform but regular meshes (in the sense of definition 6.3.11), and in this case,  $h$  is the maximum distance between two points:  $h = \max_{0 \leq j \leq n} (x_{j+1} - x_j)$ . •

**Remark 6.2.9** We can make an analogy between the convergence of a finite element method and the convergence of a finite difference method. Recall that, from the Lax theorem 2.2.20, the convergence of a finite difference scheme follows from its stability and its consistency. Let us indicate what the (formal) equivalents of these ingredients are in the context of finite elements. The role of consistency for finite elements is played by the interpolation property of lemma 6.2.5, while the role of the stability is taken by the coercivity property of the bilinear form which assures the (stable) solution of every internal approximation. •

**Proof.** Lemma 6.2.5 allows us to apply the convergence result of lemma 6.1.3 which immediately implies (6.14). To obtain (6.15), we bound the estimate of Céa's lemma 6.1.2

$$\|u - u_h\|_{H^1(0,1)} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1(0,1)} \leq C\|u - r_h u\|_{H^1(0,1)},$$

and the proof is finished thanks to lemma 6.2.5. □

We now give the proof of lemma 6.2.5 in the form of two other technical lemmas.

**Lemma 6.2.10** *There exists a constant  $C$  independent of  $h$  such that, for all  $v \in H^2(0, 1)$ ,*

$$\|v - r_h v\|_{L^2(0,1)} \leq Ch^2 \|v''\|_{L^2(0,1)}, \quad (6.16)$$

and

$$\|v' - (r_h v)'\|_{L^2(0,1)} \leq Ch \|v''\|_{L^2(0,1)}. \quad (6.17)$$

**Proof.** Take  $v \in C^\infty([0, 1])$ . By definition, the interpolant  $r_h v$  is an affine function and, for all  $x \in ]x_j, x_{j+1}[$ , we have

$$\begin{aligned} v(x) - r_h v(x) &= v(x) - \left( v(x_j) + \frac{v(x_{j+1}) - v(x_j)}{x_{j+1} - x_j} (x - x_j) \right) \\ &= \int_{x_j}^x v'(t) dt - \frac{x - x_j}{x_{j+1} - x_j} \int_{x_j}^{x_{j+1}} v'(t) dt \\ &= (x - x_j) v'(x_j + \theta_x) - (x - x_j) v'(x_j + \theta_j) \\ &= (x - x_j) \int_{x_j + \theta_j}^{x_j + \theta_x} v''(t) dt, \end{aligned} \quad (6.18)$$

by application of the finite growth formula with  $0 \leq \theta_x \leq x - x_j$  and  $0 \leq \theta_j \leq h$ . We deduce by using the Cauchy–Schwarz inequality

$$|v(x) - r_h v(x)|^2 \leq h^2 \left( \int_{x_j}^{x_{j+1}} |v''(t)| dt \right)^2 \leq h^3 \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt. \quad (6.19)$$

Integrating (6.19) with respect to  $x$  over the interval  $[x_j, x_{j+1}]$ , we obtain

$$\int_{x_j}^{x_{j+1}} |v(x) - r_h v(x)|^2 dx \leq h^4 \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt,$$

which, by summation in  $j$ , gives exactly (6.16). By density, this result is also true for all  $v \in H^2(0, 1)$ . The proof of (6.17) is similar: for  $v \in C^\infty([0, 1])$  and  $x \in ]x_j, x_{j+1}[$  we write

$$\begin{aligned} v'(x) - (r_h v)'(x) &= v'(x) - \frac{v(x_{j+1}) - v(x_j)}{h} = \frac{1}{h} \int_{x_j}^{x_{j+1}} (v'(x) - v'(t)) dt \\ &= \frac{1}{h} \int_{x_j}^{x_{j+1}} \int_t^x v''(y) dy dt. \end{aligned}$$

Squaring this inequality, applying Cauchy–Schwarz twice, and summing in  $j$  we obtain (6.17), which is also valid for all  $v \in H^2(0, 1)$  by density.  $\square$

**Lemma 6.2.11** *There exists a constant  $C$  independent of  $h$  such that, for all  $v \in H^1(0, 1)$ ,*

$$\|r_h v\|_{H^1(0,1)} \leq C \|v\|_{H^1(0,1)}, \quad (6.20)$$

and

$$\|v - r_h v\|_{L^2(0,1)} \leq Ch \|v'\|_{L^2(0,1)}. \quad (6.21)$$

Moreover, for all  $v \in H^1(0,1)$ , we have

$$\lim_{h \rightarrow 0} \|v' - (r_h v)'\|_{L^2(0,1)} = 0. \quad (6.22)$$

**Proof.** The proofs of (6.20) and (6.21) are in the same spirit as those of the preceding lemma. Take  $v \in H^1(0,1)$ . First of all we have,

$$\|r_h v\|_{L^2(0,1)} \leq \max_{x \in [0,1]} |r_h v(x)| \leq \max_{x \in [0,1]} |v(x)| \leq C \|v\|_{H^1(0,1)},$$

by virtue of lemma 4.3.3. On the other hand, since  $r_h v$  is affine, and thanks to the property (4.8) of lemma 4.3.3 which confirms that  $v$  is the primitive of  $v'$ , we have

$$\begin{aligned} \int_{x_j}^{x_{j+1}} |(r_h v)'(x)|^2 dx &= \frac{(v(x_{j+1}) - v(x_j))^2}{h} \\ &= \frac{1}{h} \left( \int_{x_j}^{x_{j+1}} v'(x) dx \right)^2 \leq \int_{x_j}^{x_{j+1}} |v'(x)|^2 dx, \end{aligned}$$

by Cauchy–Schwarz, this, by summation in  $j$ , leads to (6.20). To obtain (6.21) we take the second equation of (6.18) from which we deduce

$$|v(x) - r_h v(x)| \leq 2 \int_{x_j}^{x_{j+1}} |v'(t)| dt.$$

By squaring, using Cauchy–Schwarz, integrating with respect to  $x$ , then summing over  $j$ , we obtain (6.21).

Let us move on to the proof of (6.22). Take  $\epsilon > 0$ . As  $C^\infty([0,1])$  is dense in  $H^1(0,1)$ , for all  $v \in H^1(0,1)$  there exists  $\phi \in C^\infty([0,1])$  such that

$$\|v' - \phi'\|_{L^2(0,1)} \leq \epsilon.$$

Now  $r_h$  is a linear mapping which satisfies (6.20), therefore we deduce

$$\|(r_h v)' - (r_h \phi)'\|_{L^2(0,1)} \leq C \|v' - \phi'\|_{L^2(0,1)} \leq C\epsilon.$$

The choice of  $\phi$  and of  $\epsilon$  being fixed, we deduce from (6.17) applied to  $\phi$  that, for  $h$  sufficiently small,

$$\|\phi' - (r_h \phi)'\|_{L^2(0,1)} \leq \epsilon.$$

Consequently, summing these last three inequalities we obtain

$$\|v' - (r_h v)'\|_{L^2(0,1)} \leq \|v' - \phi'\|_{L^2} + \|\phi' - (r_h \phi)'\|_{L^2} + \|(r_h v)' - (r_h \phi)'\|_{L^2} \leq C\epsilon,$$

which implies (6.22).  $\square$

**Exercise 6.2.5** Prove the equivalent of the convergence theorem 6.2.6 for the Neumann problem (6.13).

### 6.2.3 $\mathbb{P}_2$ finite elements

The  $\mathbb{P}_2$  finite element method uses the discrete space

$$V_h = \{v \in C([0, 1]) \text{ such that } v|_{[x_j, x_{j+1}]} \in \mathbb{P}_2 \text{ for all } 0 \leq j \leq n\}, \quad (6.23)$$

and its subspace

$$V_{0h} = \{v \in V_h \text{ such that } v(0) = v(1) = 0\}. \quad (6.24)$$

The  $\mathbb{P}_2$  finite element method is the method of internal variational approximation of Section 6.1.2 applied to these spaces  $V_h$  or  $V_{0h}$ . These are composed of continuous, piecewise quadratic functions that we can represent with the help of very simple basis functions.

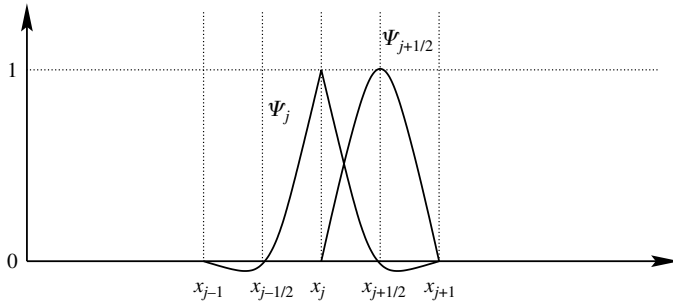


Figure 6.3. Basis functions for  $\mathbb{P}_2$  finite elements.

Let us introduce first of all the midpoints of the segments  $[x_j, x_{j+1}]$  defined by  $x_{j+1/2} = x_j + h/2$  for  $0 \leq j \leq n$ . We also define two ‘reference’ functions

$$\phi(x) = \begin{cases} (1+x)(1+2x) & \text{if } -1 \leq x \leq 0, \\ (1-x)(1-2x) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{if } |x| > 1, \end{cases}$$

and

$$\psi(x) = \begin{cases} 1 - 4x^2 & \text{if } |x| \leq 1/2, \\ 0 & \text{if } |x| > 1/2. \end{cases}$$

If the mesh is uniform, for  $0 \leq j \leq n+1$  we define the basis functions (see Figure 6.3)

$$\psi_j(x) = \phi\left(\frac{x-x_j}{h}\right), \quad 0 \leq j \leq n+1, \quad \text{and} \quad \psi_{j+1/2}(x) = \psi\left(\frac{x-x_{j+1/2}}{h}\right), \quad 0 \leq j \leq n.$$

**Lemma 6.2.12** *The space  $V_h$ , defined by (6.23), is a subspace of  $H^1(0, 1)$  of dimension  $2n+3$ , and every function  $v_h \in V_h$  is defined uniquely by its values at the vertices*

$(x_j)_{0 \leq j \leq n+1}$  and at the midpoints  $(x_{j+1/2})_{0 \leq j \leq n}$

$$v_h(x) = \sum_{j=0}^{n+1} v_h(x_j) \psi_j(x) + \sum_{j=0}^n v_h(x_{j+1/2}) \psi_{j+1/2}(x) \quad \forall x \in [0, 1].$$

Likewise,  $V_{0h}$ , defined by (6.24), is a subspace of  $H_0^1(0, 1)$  of dimension  $2n + 1$ , and every function  $v_h \in V_{0h}$  is defined uniquely by its values at the vertices  $(x_j)_{1 \leq j \leq n}$  and at the midpoints  $(x_{j+1/2})_{0 \leq j \leq n}$

$$v_h(x) = \sum_{j=1}^n v_h(x_j) \phi_j(x) + \sum_{j=0}^n v_h(x_{j+1/2}) \psi_{j+1/2}(x) \quad \forall x \in [0, 1].$$

**Remark 6.2.13** Here again,  $V_h$  is a space of Lagrange finite elements (cf. remark 6.2.2). As the functions are locally  $\mathbb{P}_2$ , we say that the space  $V_h$ , defined by (6.23), is the space of the Lagrange finite elements of order 2. •

**Proof.** By application of lemma 4.3.19  $V_h$  and  $V_{0h}$  are subspaces of  $H^1(0, 1)$ . Their dimension and the proposed bases are easily found by remarking that  $\psi_j(x_i) = \delta_{ij}$ ,  $\psi_{j+1/2}(x_{i+1/2}) = \delta_{ij}$ ,  $\psi_j(x_{i+1/2}) = 0$ ,  $\psi_{j+1/2}(x_i) = 0$  (see Figure 6.3). □

Let us now describe the **practical solution** of the Dirichlet problem (6.7) by the  $\mathbb{P}_2$  finite element method. The variational formulation (6.2) of the internal approximation reduces to solving in  $\mathbb{R}^{2n+1}$  the linear system

$$\mathcal{K}_h U_h = b_h. \quad (6.25)$$

To make this linear system explicit, it is convenient to change the indices denoting from now on the points  $(x_{1/2}, x_1, x_{3/2}, x_2, \dots, x_{n+1/2})$  in the form  $(x_{k/2})_{1 \leq k \leq 2n+1}$ , and the basis  $(\psi_{1/2}, \psi_1, \psi_{3/2}, \psi_2, \dots, \psi_{n+1/2})$  of  $V_{0h}$  in the form  $(\psi_{k/2})_{1 \leq k \leq 2n+1}$ . In this basis,  $U_h \in \mathbb{R}^{2n+1}$  is the vector of the coordinates of the approximate solution  $u_h$  which satisfies

$$u_h(x) = \sum_{k=1}^{2n+1} (U_h)_{k/2} \psi_{k/2}(x) \quad \text{with } (U_h)_{k/2} = u_h(x_{k/2}), \quad (6.26)$$

and we have

$$\mathcal{K}_h = \left( \int_0^1 \psi'_{k/2}(x) \psi'_{l/2}(x) dx \right)_{1 \leq k, l \leq 2n+1}, \quad b_h = \left( \int_0^1 f(x) \psi_{k/2}(x) dx \right)_{1 \leq k \leq 2n+1}.$$

The basis functions  $\psi_{k/2}$  have a ‘small’ support, and most of the coefficients of  $\mathcal{K}_h$  are therefore zero. A simple calculation shows that the stiffness matrix  $\mathcal{K}_h$  is here

pentadiagonal

$$\mathcal{K}_h = h^{-1} \begin{pmatrix} 16/3 & -8/3 & 0 & & & & & \\ -8/3 & 14/3 & -8/3 & 1/3 & & & & 0 \\ 0 & -8/3 & 16/3 & -8/3 & 0 & & & \\ & 1/3 & -8/3 & 14/3 & -8/3 & 1/3 & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & 0 & -8/3 & 16/3 & -8/3 & 0 \\ & 0 & & & 1/3 & -8/3 & 14/3 & -8/3 \\ & & & & & 0 & -8/3 & 16/3 \end{pmatrix}.$$

Let us remark that this matrix is more ‘full’ than that obtained by the  $\mathbb{P}_1$  finite element method, and therefore that the solution of the linear system will be more costly in calculation time. To evaluate the right-hand side  $b_h$  we use the same quadrature formulas (or numerical integration formulas) presented in the  $\mathbb{P}_1$  method.

**Theorem 6.2.14** *Let  $u \in H_0^1(0, 1)$  and  $u_h \in V_{0h}$  be the solutions of (6.7) and (6.25)–(6.26), respectively. Then, the  $\mathbb{P}_2$  finite element method converges, that is,*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(0,1)} = 0.$$

Moreover, if  $u \in H^3(0, 1)$  (which is true if  $f \in H^1(0, 1)$ ), then there exists a constant  $C$  independent of  $h$  such that

$$\|u - u_h\|_{H^1(0,1)} \leq Ch^2 \|u'''\|_{L^2(0,1)}.$$

**Exercise 6.2.6** Generalizing the preceding arguments, prove theorem 6.2.14.

Theorem 6.2.14 shows the principal advantage of  $\mathbb{P}_2$  finite elements: if the solution is regular, then the convergence of the method is **quadratic** (the rate of convergence is proportional to  $h^2$ ) while the convergence for  $\mathbb{P}_1$  finite elements is only linear (proportional to  $h$ ). Of course, this advantage has a price: there are twice as many unknowns (exactly  $2n + 1$  instead of  $n$  for  $\mathbb{P}_1$  finite elements) therefore the matrix is twice as large, also the matrix has five nonzero diagonals instead of three in the  $\mathbb{P}_1$  case. Let us remark that if the solution is not regular ( $u \in H^3(0, 1)$ ) there is no theoretical (or practical) advantage in the use of  $\mathbb{P}_2$  finite elements rather than  $\mathbb{P}_1$ .

## 6.2.4 Qualitative properties

We know that the solution of a Dirichlet problem satisfies the maximum principle (see theorem 5.2.22). It is important to know if this property is conserved by the internal variational approximation.

**Proposition 6.2.15 (discrete maximum principle)** *We assume that  $f \geq 0$  almost everywhere in  $]0, 1[$ . Then, the solution  $u_h$  of the variational approximation (6.11) by the  $\mathbb{P}_1$  finite element method satisfies  $u_h \geq 0$  in  $[0, 1]$ .*

**Proof.** Let  $u_h$  be the solution of (6.11). From lemma 6.2.1, we have

$$u_h(x) = \sum_{j=1}^n u_h(x_j) \phi_j(x),$$

where the functions  $\phi_j$  are the basis functions of the  $\mathbb{P}_1$  finite elements in  $V_{0h}$  and  $U_h = (u_h(x_j))_{1 \leq j \leq n}$  is the solution of the linear system

$$\mathcal{K}_h U_h = b_h. \quad (6.27)$$

The functions  $\phi_j$  are the ‘hat’ functions (see Figure 6.1) which are non negative: it is sufficient therefore to show that all the components of the vector  $U_h = (U_h^j)_{1 \leq j \leq n}$  are non negative to prove that the function  $u_h$  is non negative on  $[0, 1]$ . Recall that, setting  $U_h^0 = U_h^{n+1} = 0$ , the linear system (6.27) is equivalent to

$$-U_h^{j-1} + 2U_h^j - U_h^{j+1} = h b_h^j \quad \text{for all } 1 \leq j \leq n. \quad (6.28)$$

Let  $U_h^{j_0} = \min_j U_h^j$  be the smallest component of  $U_h$ : if there are several small components, we choose the one with the smallest index  $j_0$ . If  $j_0 = 0$ , then  $U_h^j \geq U_h^0 = 0$  for all  $j$ , which is the result sought. If  $j_0 \geq 1$ , then  $U_h^{j_0} < U_h^0 = 0$ , and as  $U_h^{n+1} = 0$  we deduce that  $j_0 \leq n$ . Since  $b_h^j = \int_0^1 f \psi_j dx \geq 0$  by the hypothesis on  $f$ , we can deduce the relation (6.28) for  $j_0$

$$\left( U_h^{j_0} - U_h^{j_0-1} \right) + \left( U_h^{j_0} - U_h^{j_0+1} \right) \geq 0,$$

which is a contradiction with the (strict) minimal character of  $U_h^{j_0}$ . Consequently, the  $\mathbb{P}_1$  finite element method satisfies the discrete maximum principle.  $\square$

We have proved in the preceding sections the theoretical convergence results. We can **verify the predicted rates of convergence numerically** by solving the Dirichlet problem (6.7) by the finite element with meshes of distinct size. Let us consider the following example

$$\begin{cases} -((1+x)u')' + (1 + \cos(\pi x)) u = f & \text{for } 0 < x < 1 \\ u(0) = u(1) = 0 \end{cases} \quad (6.29)$$

with  $f(x) = -\pi \cos(\pi x) + \sin(\pi x)(1 + \cos(\pi x) + \pi^2(1+x))$ , whose exact solution is  $u(x) = \sin(\pi x)$ . The ideal would be to calculate the exact error  $\|u - u_h\|_{H^1(0,1)}$ , but this needs us to make a precise calculation of the integrals, which is not easy if the solution  $u$  is complicated. In practice (and this is what we do here) we are happy to calculate the error projected into  $V_h$ , that is, we calculate  $\|r_h(u - u_h)\|_{H^1(0,1)}$  (we also say that we use the **discrete norm** in  $V_h$ ). The interest of this approach is that we can exactly calculate the integrals since  $r_h(u - u_h) = r_h u - u_h \in V_h$  (this reduces to not taking account of the interpolation errors between  $H^1(0,1)$  and  $V_h$ ). We draw

this discrete error  $\|r_h(u - u_h)\|_{H^1(0,1)}$  as a function of the mesh size  $h$ . When the solution is regular, theorem 6.2.6 predicts linear convergence (in  $h$ ) of the error in the  $\mathbb{P}_1$  finite element method, while theorem 6.2.14 predicts quadratic convergence (in  $h^2$ ) of the error in the  $\mathbb{P}_2$  finite element method. In the case of example (6.29) we draw this error for different values of  $h$  in Figure 6.4 (in logarithmic scale). The crosses and the circles correspond to the results of the calculation, the lines are the lines of reference corresponding to the functions  $h^2$  and  $h^3$ , respectively. Let us remark that the use of a logarithmic scale allows us to visualize the rates of convergence as the slope of the logarithm of the error as a function of the logarithm of  $h$ . We observe therefore a phenomenon of **superconvergence**, that is, the finite elements converge more rapidly than was proved by the theory: the error is  $h^2$  for the  $\mathbb{P}_1$  method and  $h^3$  for the  $\mathbb{P}_2$ . This gain is due to the uniformity of the mesh and to the choice of the discrete norm in  $V_h$ .

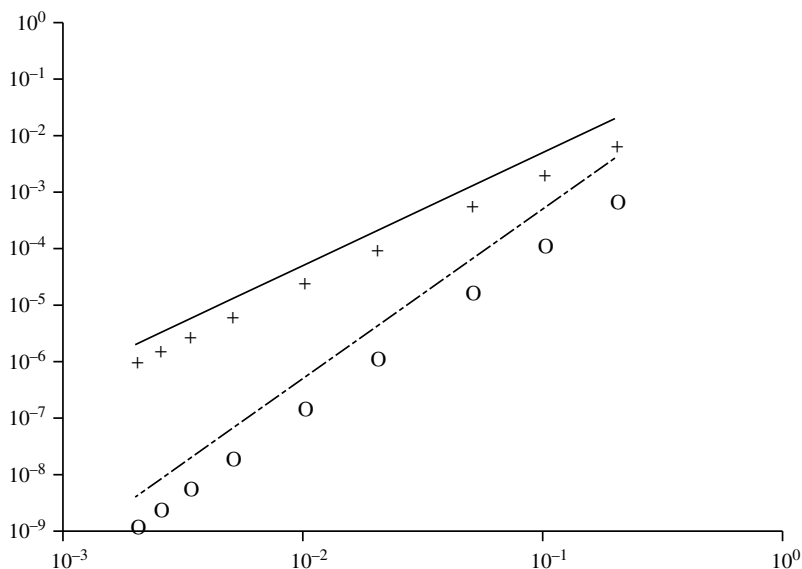


Figure 6.4. Case of a regular solution: example (6.29). Discrete  $H^1$  norm of the error as a function of the size  $h$  of the mesh (the crosses correspond to  $\mathbb{P}_1$  finite elements, the circles to  $\mathbb{P}_2$  finite elements, the lines are the graphs of  $h \rightarrow h^2$  and  $h \rightarrow h^3$ ).

If the solution is not regular, we always have convergence but with a weaker rate than is predicted in the regular case by the theorems 6.2.6 and 6.2.14. To obtain a nonregular solution, we take a right-hand side in  $H^{-1}(0,1)$  which does not belong to  $L^2(0,1)$ . In one space dimension we can therefore take a Dirac mass. Let us consider therefore the example

$$\begin{cases} -u'' = 6x - 2 + \delta_{1/2} & \text{for } 0 < x < 1 \\ u(0) = u(1) = 0 \end{cases} \quad (6.30)$$



with  $\delta_{1/2}$  the Dirac mass at the point  $x = 1/2$ , whose exact solution is  $u(x) = 1/2 - |x - 1/2| + x^2(1 - x)$ .

We draw the error  $\|r_h(u - u_h)\|_{H^1(0,1)}$  as a function of  $h$  in Figure 6.5 (in logarithmic scale). The crosses and the circles correspond to the results of the calculation ( $\mathbb{P}_1$  or  $\mathbb{P}_2$  respectively), the lines are of the lines of reference corresponding to the functions  $\sqrt{h}$  and  $h$  respectively. We see that the  $\mathbb{P}_1$  and  $\mathbb{P}_2$  finite elements converge at the same rate proportional to  $\sqrt{h}$ , which is less than the rate of  $h$  (or even  $h^2$ ) predicted in the regular case. In particular, there is no interest in using  $\mathbb{P}_2$  finite elements, rather than  $\mathbb{P}_1$ , in such a case.

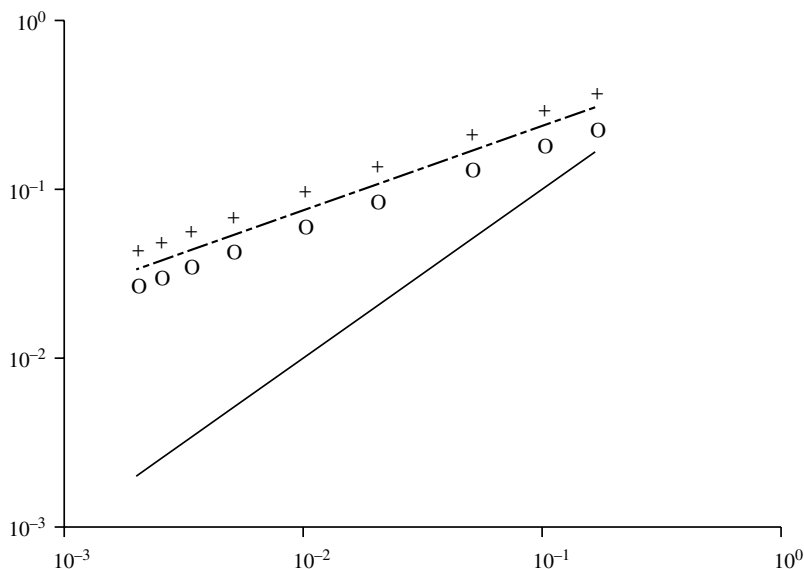


Figure 6.5. Case of a nonregular solution: example (6.30). Discrete  $H^1$  norm of the error as a function of the mesh size  $h$  (the crosses correspond to  $\mathbb{P}_1$  finite elements, the circles to  $\mathbb{P}_2$  finite elements, the lines are the graphs of  $h \rightarrow \sqrt{h}$  and  $h \rightarrow h$ ).

To calculate the error in Figures 6.4 and 6.5 we have used an exact solution. However, if this is not known, we can replace it by the approximate solution obtained with the finest mesh (assumed to be the most ‘converged’). This **numerical convergence** procedure can also be implemented for those methods for which we do not have a convergence theorem. This is often the only ‘heuristic’ way to verify if an algorithm converges and at what rate with respect to the refinement of the mesh.

### 6.2.5 Hermite finite elements

After having defined  $\mathbb{P}_1$  and  $\mathbb{P}_2$  finite elements, the reader can easily imagine how to generalize and define  $\mathbb{P}_k$  finite elements with  $k \in \mathbb{N}^*$ . These finite elements, called

Lagrange, use basis functions which are only continuous but not continuously differentiable. However, it is clear that the polynomials in  $\mathbb{P}_3$  can be connected in a continuously differentiable way. In this case the values of the derivatives will also be used to characterize the functions (see remark 6.2.2). We therefore introduce the method of  $\mathbb{P}_3$  **Hermite finite elements** which are defined on the discrete space

$$V_h = \{v \in C^1([0, 1]) \text{ such that } v|_{[x_j, x_{j+1}]} \in \mathbb{P}_3 \text{ for all } 0 \leq j \leq n\}. \quad (6.31)$$

We must pay attention to the fact that in the definition (6.31) of  $V_h$  we ask for the functions to belong to  $C^1([0, 1])$ , and not only to  $C([0, 1])$ . This is the difference between Hermite and Lagrange finite elements respectively.

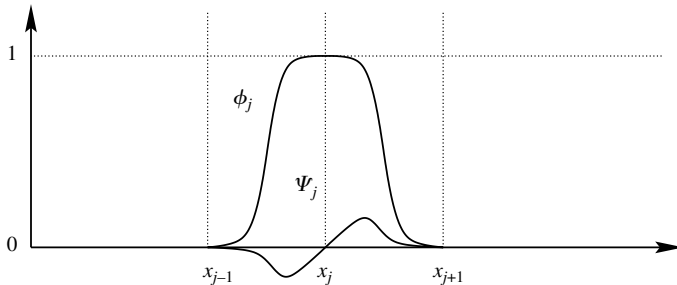


Figure 6.6. The basis functions of  $\mathbb{P}_3$  Hermite finite elements.

We can represent the functions of  $V_h$  with the help of very simple basis functions. We define two ‘reference’ functions

$$\phi(x) = \begin{cases} (1+x)^2(1-2x) & \text{if } -1 \leq x \leq 0, \\ (1-x)^2(1+2x) & \text{if } 0 \leq x \leq 1, \\ 0 & \text{if } |x| > 1, \end{cases}$$

and

$$\psi(x) = \begin{cases} x(1+x)^2 & \text{if } -1 \leq x \leq 0, \\ x(1-x)^2 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{if } |x| > 1. \end{cases}$$

If the mesh is uniform, for  $0 \leq j \leq n+1$  we define the basis functions (see Figure 6.6)

$$\phi_j(x) = \phi\left(\frac{x-x_j}{h}\right) \text{ for } 0 \leq j \leq n+1, \quad \psi_j(x) = \psi\left(\frac{x-x_j}{h}\right) \text{ for } 0 \leq j \leq n+1.$$

**Lemma 6.2.16** *The space  $V_h$ , defined by (6.31), is a subspace of  $H^1(0, 1)$  of dimension  $2(n+2)$ . Every function  $v_h$  of  $V_h$  is defined uniquely by **its values and those of its derivative** at the vertices  $(x_j)_{0 \leq j \leq n+1}$ , and we have for all  $x \in [0, 1]$*

$$v_h(x) = \sum_{j=0}^{n+1} v_h(x_j) \phi_j(x) + \sum_{j=0}^{n+1} (v_h)'(x_j) \psi_j(x). \quad (6.32)$$

**Proof.** The functions of  $V_h$  being of class  $C^1$ , it is a subspace of  $H^1(0, 1)$ . We verify easily that the  $(\phi_j, \psi_j)$  form a basis of  $V_h$  by remarking that  $\phi_j(x_i) = \delta_{ij}$ ,  $\psi_j(x_i) = 0$ ,  $\phi'_j(x_i) = 0$ ,  $\psi'_j(x_i) = \delta_{ij}$  (see Figure 6.6).  $\square$

We can use the space  $V_h$  (or at least its subspace of functions which are zero at 0 and 1) to solve the Dirichlet problem (6.7), but this is not the most current use of  $V_h$ . In practice we use  $V_h$  to solve the **plate equation** (see (5.70) and Chapter 1), or beams in  $N = 1$  dimension,

$$\begin{cases} u'''' = f & \text{in } ]0, 1[ \\ u(0) = u(1) = u'(0) = u'(1) = 0, \end{cases} \quad (6.33)$$

which has a unique solution  $u \in H_0^2(0, 1)$  if  $f \in L^2(0, 1)$ . In effect,  $V_h$  is not only a subspace of  $H^1(0, 1)$ , but is also a subspace of  $H^2(0, 1)$  (this is not the case for Lagrange finite elements). To solve (6.33) we shall need the subspace

$$V_{0h} = \{v \in V_h \text{ such that } v(0) = v(1) = v'(0) = v'(1) = 0\}. \quad (6.34)$$

**Lemma 6.2.17** *The space  $V_h$ , and its subspace  $V_{0h}$  defined by (6.34), are subspaces of  $H^2(0, 1)$ , and of  $H_0^2(0, 1)$  respectively, of dimensions  $2(n+2)$ , and  $2n$  respectively. Every function  $v_h$  of  $V_{0h}$  is defined uniquely by its values and those of its derivative at the vertices  $(x_j)_{1 \leq j \leq n}$ , and we have for all  $x \in [0, 1]$*

$$v_h(x) = \sum_{j=1}^n v_h(x_j) \phi_j(x) + \sum_{j=1}^n (v_h)'(x_j) \psi_j(x).$$

**Proof.** Take  $v_h \in V_h$ : it is of class  $C^1$  over  $[0, 1]$  and piecewise  $C^2$ . Therefore, its derivative  $v'_h$ , being continuous and piecewise  $C^1$ , belongs to  $H^1(0, 1)$  (by virtue of lemma 4.3.19). Consequently,  $v_h$  is an element of  $H^2(0, 1)$ . The rest of the lemma is similar to lemma 6.2.16  $\square$

Let us now describe briefly the practical solution of the plate equation (6.33) by the  $\mathbb{P}_3$  Hermite finite element method. The variational formulation of the internal approximation is

$$\text{find } u_h \in V_{0h} \text{ such that } \int_0^1 u_h''(x) v_h''(x) dx = \int_0^1 f(x) v_h(x) dx \quad \forall v_h \in V_{0h}. \quad (6.35)$$

We decompose  $u_h$  in the basis of  $(\phi_j, \psi_j)_{1 \leq j \leq n}$  and we denote by  $U_h = (u_h(x_j), u'_h(x_j))_{1 \leq j \leq n}$  the vector of its coordinates in this basis. The variational formulation (6.35) reduces to solving in  $\mathbb{R}^{2n}$  a linear system

$$\mathcal{K}_h U_h = b_h.$$

**Exercise 6.2.7** Explicitly calculate the stiffness matrix  $\mathcal{K}_h$  for (6.35).

## 6.3 Finite elements in $N \geq 2$ dimensions

We now place ourselves in  $N \geq 2$  space dimensions (in practice  $N = 2, 3$ ). To simplify the exposition, certain results will only be proved in  $N = 2$  dimensions, but they extend to  $N = 3$  dimensions (with the cost, sometimes, of important technical and practical complications).

We consider the model Dirichlet problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (6.36)$$

which we know has a unique solution in  $H_0^1(\Omega)$ , if  $f \in L^2(\Omega)$  (see Chapter 5).

In all that follows we shall assume that the domain  $\Omega$  is a **polyhedral** (polygonal if  $N = 2$ ), that is,  $\overline{\Omega}$  is a finite union of polyhedra of  $\mathbb{R}^N$ . Let us recall that a polyhedron is a finite intersection of half-spaces of  $\mathbb{R}^N$  and that the parts of its boundary which belong to a single hyperplane are called its faces. The reason for this hypothesis is that it is only possible to mesh exactly such open sets. We shall describe later what happens for general domains bounded by ‘curves’ (see remark 6.3.18).

### 6.3.1 Triangular finite elements

We start with the definition of a mesh of the domain  $\Omega$  by triangles in  $N = 2$  dimensions and by tetrahedra in  $N = 3$  dimensions. We group the triangles and the tetrahedra in the more general family of  $N$ -simplices. We call the  $N$ -simplex  $K$  of  $\mathbb{R}^N$  the convex envelope of  $(N + 1)$  points  $(a_j)_{1 \leq j \leq N+1}$  of  $\mathbb{R}^N$ , called the vertices of  $K$ . Of course, a 2-simplex is simply a triangle and a 3-simplex a tetrahedron (see Figure 6.9). We say that the  $N$ -simplex  $K$  is nondegenerate if the points  $(a_j)_{1 \leq j \leq N+1}$  do not belong to the same hyperplane of  $\mathbb{R}^N$  (the triangle or the tetrahedron is not ‘flat’). If we denote by  $(a_{i,j})_{1 \leq i \leq N}$  the coordinates of the vector  $a_j$ , the nondegeneracy condition on  $K$  is that the matrix

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N+1} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N+1} \\ \vdots & \vdots & & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N+1} \\ 1 & 1 & \cdots & 1 \end{pmatrix} \quad (6.37)$$

is invertible (which we shall always assume in what follows). An  $N$ -simplex has as many faces as vertices, which are themselves  $(N - 1)$ -simplices.

**Definition 6.3.1** *Let  $\Omega$  be an open connected polyhedron of  $\mathbb{R}^N$ . A triangular mesh or a triangulation of  $\Omega$  is a set  $\mathcal{T}_h$  of (nondegenerate)  $N$ -simplices  $(K_i)_{1 \leq i \leq n}$  which satisfies*

1.  $K_i \subset \overline{\Omega}$  and  $\overline{\Omega} = \cup_{i=1}^n K_i$ .

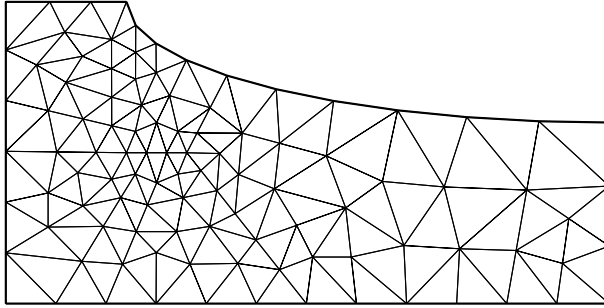


Figure 6.7. Example of triangular mesh in  $N = 2$  dimensions.

2. The intersection  $K_i \cap K_j$  of two distinct  $N$ -simplices is an  $m$ -simplex, with  $0 \leq m \leq N - 1$ , whose vertices are also vertices of  $K_i$  and  $K_j$ . (In  $N = 2$  dimensions, the intersection of two triangles is either empty, or reduced to a common vertex, or an **entire** common edge; in  $N = 3$  dimensions, the intersection of two tetrahedra is empty, or a common vertex, or an entire common edge, or an entire common face.)

The **vertices** or **nodes** of the mesh  $\mathcal{T}_h$  are the vertices of these  $N$ -simplices  $K_i$  which compose it. By convention, the parameter  $h$  denotes the maximum diameter of the  $N$ -simplices  $K_i$ .

It is clear that definition 6.3.1 can only be applied to a polyhedral open set and not to an arbitrary open set. Definition 6.3.1 contains a certain number of restrictions on the mesh: in this case, we often talk of a **conforming mesh**. An example of a conforming mesh is given in Figure 6.7, while Figure 6.8 gives some of the situations forbidden by definition 6.3.1.

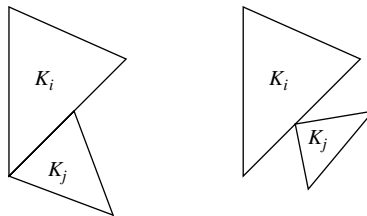


Figure 6.8. Examples of forbidden situations for a triangular mesh.

**Remark 6.3.2** We shall not say anything about the algorithms for constructing a triangular mesh. We shall be content with saying that, if it is relatively easy to mesh plane domains (there are many free pieces of software which allow us to do this), it is complicated to mesh three-dimensional domains. We refer the interested reader to the work [20] for this subject. •

**Exercise 6.3.1** Let  $\mathcal{T}_h$  be a mesh of  $\bar{\Omega}$  for  $\Omega$  a simply connected open polygonal set of  $\mathbb{R}^2$ . We denote by  $n_t$  the number of triangles of  $\mathcal{T}_h$ ,  $n_c$  the number of faces or sides of the triangles (a common side of two triangles is only counted once),  $n_s$  the number of vertices of the mesh, and  $n_{0s}$  the number of interior vertices of the mesh (which are not on  $\partial\Omega$ ). Prove the Euler relations,  $n_t + n_s = n_c + 1$  and  $3n_t + n_s = 2n_c + n_{0s}$ .

In an  $N$ -simplex  $K$  it is easy to use barycentric coordinates instead of the usual Cartesian coordinates. Recall that, if  $K$  is a nondegenerate  $N$ -simplex with vertices  $(a_j)_{1 \leq j \leq N+1}$ , the **barycentric coordinates**  $(\lambda_j)_{1 \leq j \leq N+1}$  of  $x \in \mathbb{R}^N$  are defined by

$$\sum_{j=1}^{N+1} \lambda_j = 1, \quad \sum_{j=1}^{N+1} a_{i,j} \lambda_j = x_i, \quad \text{for } 1 \leq i \leq N, \quad (6.38)$$

which have a unique solution because the matrix  $A$ , defined by (6.37), is invertible. Let us remark that  $\lambda_j$  are affine functions of  $x$ . We then verify that

$$K = \{x \in \mathbb{R}^N \text{ such that } \lambda_j(x) \geq 0 \text{ for } 1 \leq j \leq N+1\},$$

and that the  $(N+1)$  faces of  $K$  are the intersections of  $K$  and the hyperplanes  $\lambda_j(x) = 0$ ,  $1 \leq j \leq N+1$ . We can then define a set of points of  $K$  which will play a particular role in what follows: for every integer  $k \geq 1$  we call the **lattice of order  $k$**  the set

$$\sum_k = \left\{ x \in K \text{ such that } \lambda_j(x) \in \left\{ 0, \frac{1}{k}, \dots, \frac{k-1}{k}, 1 \right\} \text{ for } 1 \leq j \leq N \right\}. \quad (6.39)$$

For  $k=1$   $\sum_1$  is the set of vertices of  $K$ , and for  $k=2$   $\sum_2$  is the set of the vertices and the midpoints of the edges linking two nodes (see Figure 6.9). In the general case,  $\sum_k$  is a finite set of points  $(\sigma_j)_{1 \leq j \leq n_k}$ .

We now define the set  $\mathbb{P}_k$  of polynomials with real coefficients from  $\mathbb{R}^N$  into  $\mathbb{R}$  of degree less than or equal to  $k$ , that is, all  $p \in \mathbb{P}_k$  are written in the form

$$p(x) = \sum_{\substack{i_1, \dots, i_N \geq 0 \\ i_1 + \dots + i_N \leq k}} \alpha_{i_1, \dots, i_N} x_1^{i_1} \cdots x_N^{i_N} \text{ with } x = (x_1, \dots, x_N).$$

The interest in the idea of a lattice  $\sum_k$  of an  $N$ -simplex  $K$  is that it allows us to characterize all the polynomials of  $\mathbb{P}_k$  (we say that  $\sum_k$  is **unisolvant** for  $\mathbb{P}_k$ ).

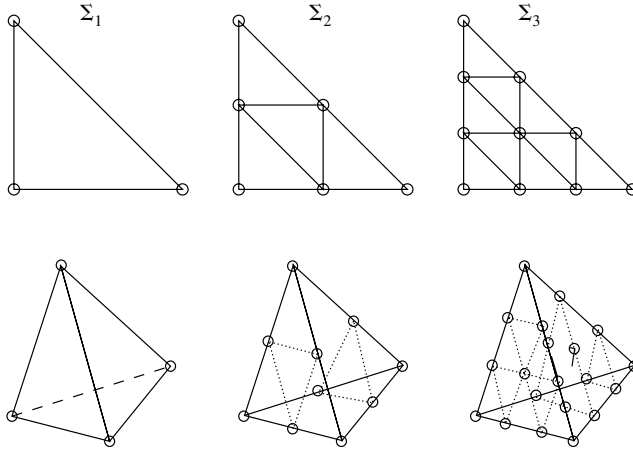


Figure 6.9. Lattice of order 1, 2, and 3 for a triangle (top) and a tetrahedron (bottom). The circles represent the points of the lattice.

**Lemma 6.3.3** *Let  $K$  be an  $N$ -simplex. For an integer  $k \geq 1$ , let  $\Sigma_k$  be the lattice of order  $k$ , defined by (6.39), whose points are denoted  $(\sigma_j)_{1 \leq j \leq n_k}$ . Then, every polynomial of  $\mathbb{P}_k$  is uniquely determined by its values at the points  $(\sigma_j)_{1 \leq j \leq n_k}$ . In other words, there exists a basis  $(\psi_j)_{1 \leq j \leq n_k}$  of  $\mathbb{P}_k$  such that*

$$\psi_j(\sigma_i) = \delta_{ij} \quad 1 \leq i, j \leq n_k.$$

**Proof.** The cardinality of  $\Sigma_k$  and the dimension of  $\mathbb{P}_k$  coincide

$$\text{card}(\Sigma_k) = \dim(\mathbb{P}_k) = \frac{(N+k)!}{N!k!}$$

(the proof is left as an exercise, at least for  $k = 1, 2$ ). Since the mapping which takes every polynomial of  $\mathbb{P}_k$  into its values on the lattice  $\Sigma_k$  is linear, it is sufficient to show that it is injective to show that it is bijective. Take therefore a polynomial  $p \in \mathbb{P}_k$  which is zero on  $\Sigma_k$ . We show, by induction on the dimension  $N$ , that  $p$  is identically zero on  $\mathbb{R}^N$ . For  $N = 1$ , it is clear that a polynomial of degree  $k$  which is zero at  $(k+1)$  distinct points is zero. Let us assume the result is true to order  $N-1$ . As  $x$  depends linearly on the barycentric coordinates  $(\lambda_j(x))_{1 \leq j \leq N+1}$ , we can define a polynomial  $q(\lambda) = p(x)$  of degree greater than  $k$  in the variable  $\lambda \in \mathbb{R}^{N+1}$ . If we fix a coordinate  $\lambda_j$  in the set  $\{0, 1/k, \dots, (k-1)/k, 1\}$  and we set  $\lambda = (\lambda', \lambda_j)$ , we obtain a polynomial  $q_j(\lambda') = q(\lambda)$  which depends on  $N-1$  independent variables (because we have the relation  $\sum_{j=1}^{N+1} \lambda_j = 1$ ) and which is zero on the section of the lattice  $\Sigma_k$  corresponding to the fixed value of  $\lambda_j$ . As this section is also the lattice of order  $k$  of a

$(N-1)$ -simplex in the fixed hyperplane  $\lambda_j$ , we can apply the induction hypothesis and deduce that  $q_j = 0$ . In other words, the factor  $\lambda_j(\lambda_j - 1/k) \cdots (\lambda_j - (k-1)/k)(\lambda_j - 1)$  divides  $q$ , which is a contradiction with the fact that the degree of  $q(\lambda)$  is less than or equal to  $k$ , except if  $q = 0$ , which is the result sought.  $\square$

**Lemma 6.3.4** *Let  $K$  and  $K'$  be two  $N$ -simplices having a common face  $\Gamma = \partial K \cap \partial K'$ . Take an integer  $k \geq 1$ . Then, their lattices of order  $k$ ,  $\Sigma_k$ , and  $\Sigma'_k$  coincide on this face  $\Gamma$ . Moreover, given  $p_K$  and  $p_{K'}$  two polynomials of  $\mathbb{P}_k$ , the function  $v$  defined by*

$$v(x) = \begin{cases} p_K(x) & \text{if } x \in K \\ p_{K'}(x) & \text{if } x \in K' \end{cases}$$

*is continuous on  $K \cup K'$ , if and only if  $p_K$  and  $p_{K'}$  have values which coincide at the points of the lattice on the common face  $\Gamma$ .*

**Proof.** It is clear that the restriction to a face of  $K$  of its lattice of order  $\Sigma_k$  is also a lattice of order  $k$  in the hyperplane containing this face, which only depends on the vertices of this face. Consequently, the lattices  $\Sigma_k$  and  $\Sigma'_k$  coincide on their common face  $\Gamma$ . If the polynomials  $p_K$  and  $p_{K'}$  coincide at the points of  $\Sigma_k \cap \Gamma$ , then by application of lemma 6.3.3 they are equal on  $\Gamma$ , which proves the continuity of  $v$ .  $\square$

In practice, we mostly use the polynomials of degree 1 or 2. In this case we have the following characterizations of  $\mathbb{P}_1$  and  $\mathbb{P}_2$  in an  $N$ -simplex  $K$ .

**Exercise 6.3.2** Let  $K$  be an  $N$ -simplex with vertices  $(a_j)_{1 \leq j \leq N+1}$ . Show that every polynomial  $p \in \mathbb{P}_1$  is in the form

$$p(x) = \sum_{j=1}^{N+1} p(a_j) \lambda_j(x),$$

where the  $(\lambda_j(x))_{1 \leq j \leq N+1}$  are the barycentric coordinates of  $x \in \mathbb{R}^N$ .

**Exercise 6.3.3** Let  $K$  be an  $N$ -simplex with vertices  $(a_j)_{1 \leq j \leq N+1}$ . We define the midpoints  $(a_{jj'})_{1 \leq j < j' \leq N+1}$  of the edges of  $K$  by their barycentric coordinates

$$\lambda_j(a_{jj'}) = \lambda_{j'}(a_{jj'}) = \frac{1}{2}, \quad \lambda_l(a_{jj'}) = 0 \quad \text{for } l \neq j, j'.$$

Verify that  $\Sigma_2$  is precisely composed of the vertices and of the midpoints of the edges and that every polynomial  $p \in \mathbb{P}_2$  is in the form

$$p(x) = \sum_{j=1}^{N+1} p(a_j) \lambda_j(x) (2\lambda_j(x) - 1) + \sum_{1 \leq j < j' \leq N+1} 4p(a_{jj'}) \lambda_j(x) \lambda_{j'}(x),$$

where the  $(\lambda_j(x))_{1 \leq j \leq N+1}$  are the barycentric coordinates of  $x \in \mathbb{R}^N$ .



We now have all the tools to define the  $\mathbb{P}_k$  finite element method.

**Definition 6.3.5** Given a mesh  $\mathcal{T}_h$  of an open connected polyhedral set  $\Omega$ , the  $\mathbb{P}_k$  finite element method, or **Lagrange triangular finite elements of order  $k$** , associated with this mesh, is defined by the discrete space

$$V_h = \{v \in C(\overline{\Omega}) \text{ such that } v|_{K_i} \in \mathbb{P}_k \text{ for all } K_i \in \mathcal{T}_h\}. \quad (6.40)$$

We call the **nodes of the degrees of freedom** the set of points  $(\hat{a}_i)_{1 \leq i \leq n_{dl}}$  of the lattice of order  $k$  of each of the  $N$ -simplices  $K_i \in \mathcal{T}_h$ . We only count once the points which coincide and  $n_{dl}$  is the number of degrees of freedom of the  $\mathbb{P}_k$  finite element method. We call the **degrees of freedom** of a function  $v \in V_h$  the set of the values of  $v$  at these vertices  $(\hat{a}_i)_{1 \leq i \leq n_{dl}}$ . We also define the subspace  $V_{0h}$  by

$$V_{0h} = \{v \in V_h \text{ such that } v = 0 \text{ on } \partial\Omega\}. \quad (6.41)$$

When  $k = 1$  the vertices of these degrees of freedom coincide with the vertices of the mesh. When  $k = 2$  these vertices are composed on the one hand of the vertices of the mesh and on the other hand of the midpoints of the edges linking two vertices.

**Remark 6.3.6** The name ‘Lagrange finite elements’ corresponds to the finite elements whose degrees of freedom are point values of the functions of the space  $V_h$ . We can define other types of finite elements, for example, Hermite finite elements (see Section 6.2.5) for which the degrees of freedom are the point values of the function and of its derivatives. •

**Proposition 6.3.7** The space  $V_h$ , defined by (6.40), is a subspace of  $H^1(\Omega)$  whose dimension is finite, equal to the number of degrees of freedom. Moreover, there exists a basis of  $V_h$   $(\phi_i)_{1 \leq i \leq n_{dl}}$  defined by

$$\phi_i(\hat{a}_j) = \delta_{ij} \quad 1 \leq i, j \leq n_{dl},$$

such that

$$v(x) = \sum_{i=1}^{n_{dl}} v(\hat{a}_i) \phi_i(x).$$

**Proof.** The elements of  $V_h$ , being regular over each mesh  $K_i$  and continuous over  $\overline{\Omega}$ , belong to  $H^1(\Omega)$  (see lemma 4.3.19). Thanks to lemma 6.3.4 the elements of  $V_h$  are exactly obtained by assembling on each  $K_i \in \mathcal{T}_h$  the polynomials of  $\mathbb{P}_k$  which coincide with the degrees of freedom of the faces (which proves in passing that  $V_h$  is not reduced to the constant functions). Finally, by assembling the bases  $(\psi_j)_{1 \leq j \leq n_k}$  of  $\mathbb{P}_k$  on each mesh  $K_i$  (provided by lemma 6.3.3) we obtain the stated basis  $(\phi_i)_{1 \leq i \leq n_{dl}}$  of  $V_h$ . □

**Remark 6.3.8** We obtain a similar result for the subspace  $V_{0h}$ , defined by (6.41), which is a subspace of  $H_0^1(\Omega)$  of finite dimension equal to the number of interior degrees of freedom (we do not count the vertices on the boundary  $\partial\Omega$ ). •

**Exercise 6.3.4** Let  $\mathcal{T}_h$  be a mesh of  $\bar{\Omega}$  for  $\Omega$ , a simply connected polygonal open set of  $\mathbb{R}^2$ . We denote by  $n_t$  the number of triangles of  $\mathcal{T}_h$ ,  $n_c$  the number of faces or sides of the triangles (a side common to two triangles is only counted once),  $n_s$  the number of vertices of the mesh, and  $n_{0s}$  the number of interior vertices of the mesh. Show that the dimensions of the spaces  $V_h$  and  $V_{0h}$  are

$$\dim V_h = \frac{k(k-1)}{2}n_t + kn_s - k + 1, \quad \dim V_{0h} = \frac{k(k+1)}{2}n_t - kn_s + k + 1.$$

Let us now describe the **practical solution** of the Dirichlet problem (6.36) by the  $\mathbb{P}_k$  finite element method. The variational formulation (6.2) of the internal approximation becomes:

$$\text{find } u_h \in V_{0h} \text{ such that } \int_{\Omega} \nabla u_h \cdot \nabla v_h \, dx = \int_{\Omega} f v_h \, dx \quad \forall v_h \in V_{0h}. \quad (6.42)$$

We decompose  $u_h$  in the basis of  $(\phi_j)_{1 \leq j \leq n_{dl}}$  and we take  $v_h = \phi_i$  which gives

$$\sum_{j=1}^{n_{dl}} u_h(\hat{a}_j) \int_{\Omega} \nabla \phi_j \cdot \nabla \phi_i \, dx = \int_{\Omega} f \phi_i \, dx.$$

By denoting  $U_h = (u_h(\hat{a}_j))_{1 \leq j \leq n_{dl}}$ ,  $b_h = (\int_{\Omega} f \phi_i \, dx)_{1 \leq i \leq n_{dl}}$ , and by introducing the **stiffness matrix**

$$\mathcal{K}_h = \left( \int_{\Omega} \nabla \phi_j \cdot \nabla \phi_i \, dx \right)_{1 \leq i, j \leq n_{dl}},$$

the variational formulation in  $V_{0h}$  reduces to solving in  $\mathbb{R}^{n_{dl}}$  the linear system

$$\mathcal{K}_h U_h = b_h.$$

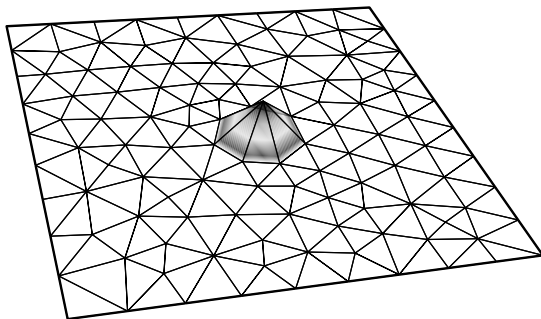
As the basis functions  $\phi_j$  have a ‘small’ support around the node  $\hat{a}_i$  (see Figure 6.10), the intersection of the supports of  $\phi_j$  and  $\phi_i$  is often empty and most of the coefficients of  $\mathcal{K}_h$  are zero. We say that the matrix  $\mathcal{K}_h$  is **sparse**.

To calculate the coefficients of  $\mathcal{K}_h$ , we can use the following exact integration formula. We denote by  $(\lambda_i(x))_{1 \leq i \leq N+1}$  the barycentric coordinates of the point under consideration  $x$  of an  $N$ -simplex  $K$ . For every  $\alpha_1, \dots, \alpha_{N+1} \in \mathbb{N}$ , we have

$$\int_K \lambda_1(x)^{\alpha_1} \cdots \lambda_{N+1}(x)^{\alpha_{N+1}} \, dx = \text{Volume}(K) \frac{\alpha_1! \cdots \alpha_{N+1}! N!}{(\alpha_1 + \cdots + \alpha_{N+1} + N)!}. \quad (6.43)$$

To calculate the right-hand side  $b_h$  (and possibly the matrix  $\mathcal{K}_h$ ), we use the **quadrature formulas** (or numerical integration formulas) which give an approximation of the integrals on each  $N$ -simplex  $K_i \in \mathcal{T}_h$ . For example, if  $K$  is an  $N$ -simplex with vertices  $(a_i)_{1 \leq i \leq N+1}$ , the following formulas generalize the ‘midpoint’ and the ‘trapezium’ formulas in one dimension:

$$\int_K \psi(x) \, dx \approx \text{Volume}(K) \psi(a_0), \quad (6.44)$$

Figure 6.10.  $\mathbb{P}_1$  basis in  $N = 2$  dimensions.

with  $a_0 = (N + 1)^{-1} \sum_{i=1}^{N+1} a_i$ , the barycentre of  $K$ , and

$$\int_K \psi(x) dx \approx \frac{\text{Volume}(K)}{N + 1} \sum_{i=1}^{N+1} \psi(a_i). \quad (6.45)$$

As was shown in exercises 6.3.6 and 6.3.8, these formulas are exact for affine functions and are therefore approximations of order 2 in  $h$  for regular functions.

The construction of the matrix  $\mathcal{K}_h$  is called **matrix assembly**. The computer implementation of this stage of the calculation can be very complicated, but its cost in terms of time of calculation is low. This is not the case for the solution of the linear system  $\mathcal{K}_h U_h = b_h$  which is the **most costly** stage of the method in calculation time (and in memory storage). In particular, the three-dimensional calculations are currently very expensive when we use fine meshes. Exercise 6.3.11 allows us to calculate this. Happily, the stiffness matrix  $\mathcal{K}_h$  is **sparse** (that is, most of its elements are zero), which allows us to minimize the calculations (for more details see the solution algorithms for linear systems in the Section 13.1 of the appendix). Recall that the matrix  $\mathcal{K}_h$  is necessarily invertible by application of lemma 6.1.1 and that it is symmetric.

**Exercise 6.3.5** Prove the formula (6.43) in  $N = 2$  dimensions.

**Exercise 6.3.6** Show that the formulas (6.44) and (6.45) are exact for  $\psi \in \mathbb{P}_1$ .

**Exercise 6.3.7** Let  $K$  be a triangle of  $\mathbb{R}^2$  with vertices  $(a_i)_{1 \leq i \leq 3}$  and with barycentre  $a_0$ . Let  $(a_{ij})_{1 \leq i < j \leq 3}$  be the midpoints of the segments with ends  $a_i, a_j$ . Show that the quadrature formula

$$\int_K \psi(x) dx \approx \frac{\text{Area}(K)}{3} \sum_{1 \leq i < j \leq 3} \psi(a_{ij})$$

is exact for  $\psi \in \mathbb{P}_2$ , while the formula

$$\int_K \psi(x) dx \approx \frac{\text{Area}(K)}{60} \left( 3 \sum_{i=1}^3 \psi(a_i) + 8 \sum_{1 \leq i < j \leq 3} \psi(a_{ij}) + 27\psi(a_0) \right)$$

is exact for  $\psi \in \mathbb{P}_3$ .

**Exercise 6.3.8** Let  $(b_i)_{1 \leq i \leq I}$  be the points of an  $N$ -simplex  $K$  and  $(\omega_i)_{1 \leq i \leq I}$  be real weights. Take a quadrature formula

$$\int_K \psi(x) dx \approx \text{Volume}(K) \sum_{i=1}^I \omega_i \psi(b_i)$$

which is exact for  $\psi \in \mathbb{P}_k$ . Show that, for a regular function  $\psi$ , we have

$$\frac{1}{\text{Volume}(K)} \int_K \psi(x) dx = \sum_{i=1}^I \omega_i \psi(b_i) + \mathcal{O}(h^{k+1}),$$

where  $h$  is the diameter of  $K$ .

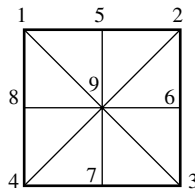


Figure 6.11. Example of mesh and of classification of the vertices.

**Exercise 6.3.9** We consider the square  $\Omega = ]-1, +1[^2$  meshed as in Figure 6.11. Calculate the stiffness matrix  $\mathcal{K}_h$  of  $\mathbb{P}_1$  finite elements applied to the Laplacian with Neumann boundary conditions (use the symmetries of the mesh).

**Exercise 6.3.10** Apply the  $\mathbb{P}_1$  finite element method to the Dirichlet problem (6.36) in the square  $\Omega = ]0, 1[^2$  with the uniform triangular mesh of Figure 6.12. Show that the stiffness matrix  $\mathcal{K}_h$  is the same matrix than that we would obtain by application of the finite difference method (up to a multiplicative factor  $h^2$ ), but that the right-hand side  $b_h$  is different.

**Exercise 6.3.11** We reuse the notation of exercise 6.3.10. We denote by  $n$  the number of points of the mesh on a side of the square (assumed to be the same for each side).

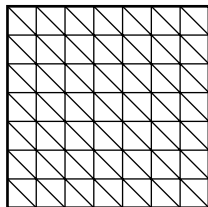


Figure 6.12. Uniform triangular mesh of a square.

We number the vertices of the mesh (or the degrees of freedom) ‘line by line’. Show that the  $\mathbb{P}_1$  finite element stiffness matrix  $\mathcal{K}_h$  has size of the order of  $n^2$  and bandwidth of the order of  $2n$  (for large  $n$ ).

Show that the same method and the same type of mesh for the cube  $\Omega = ]0, 1[^3$  leads to a matrix whose size is of the order of  $n^3$  and bandwidth of the order of  $2n^2$  (where  $n$  is the number of vertices along an edge of the cube  $\Omega$ ).

**Remark 6.3.9** As we show in the example of exercise 6.3.10, the way of numbering the vertices of these degrees of freedom (or equivalently the basis functions) has an influence on the sparse structure of the matrix  $\mathcal{K}_h$ , that is, on the position of its nonzero elements. As explained in Section 13.1 (see, for example, lemma 13.1.4), this sparse structure of the matrix has a large influence on the performance of the solution of the linear system  $\mathcal{K}_h U_h = b_h$ . For example, if we solve this linear system by a ‘Gaussian elimination’ method, it is advantageous to choose a numbering which groups the nonzero elements close to the diagonal. •

**Remark 6.3.10** To simplify the analysis (and also the implementation) we can use an affine transformation to reduce every  $N$ -simplex  $K$  of the mesh  $\mathcal{T}_h$  to a ‘reference’  $N$ -simplex  $K_0$ . By this simple change of variable, all the calculations are reduced to calculations on  $K_0$ . In practice, we often choose

$$K_0 = \left\{ x \in \mathbb{R}^N \text{ such that } \sum_{i=1}^N x_i \leq 1, x_i \geq 0 \text{ for } 1 \leq i \leq N \right\}, \quad (6.46)$$

and we see easily that every  $N$ -simplex  $K$  is the image, by an affine transformation, of  $K_0$ . In effect, the barycentric coordinates, defined by (6.38) are the same for  $K$  and  $K_0$  and, by denoting  $\lambda = (\lambda_j)_{1 \leq j \leq N+1}$ ,  $\tilde{x} = (x, 1)$  the point under consideration in  $K$ ,  $\tilde{x}_0 = (x_0, 1)$  this point in  $K_0$ , we have  $A\lambda = \tilde{x}$ , and  $A_0\lambda = \tilde{x}_0$ , where the matrices  $A$  and  $A_0$  are defined by (6.37) and invertible. We therefore deduce that  $\tilde{x} = AA_0^{-1}\tilde{x}_0$ , that is, there exists a matrix  $B$ , invertible of order  $N$ , and a vector  $b \in \mathbb{R}^N$  such that  $x = Bx_0 + b$ . •

The following exercise shows that the  $\mathbb{P}_1$  finite element method satisfies the maximum principle.

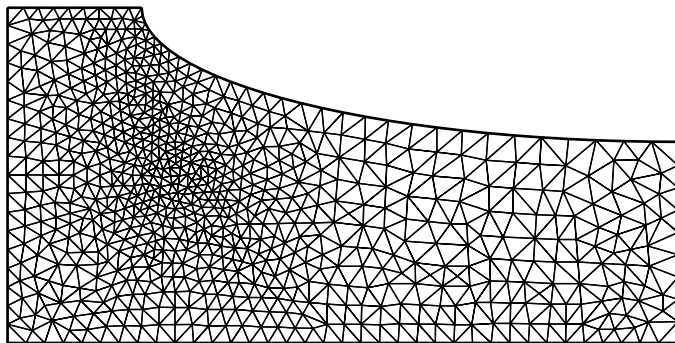


Figure 6.13. Triangular mesh finer than that of figure 6.7.

**Exercise 6.3.12** We say that a real square matrix  $B = (b_{ij})_{1 \leq i, j \leq n}$  is an M-matrix if, for all  $i$ ,

$$b_{ii} > 0, \quad \sum_{k=1}^n b_{ik} > 0, \quad b_{ij} \leq 0 \quad \forall j \neq i.$$

Show that every M-matrix is invertible and that all the coefficients of its inverse are positive or zero.

**Exercise 6.3.13** We work in  $N = 2$  dimensions. Let  $u_h$  be the approximate solution of the Dirichlet problem (6.36) obtained by the  $\mathbb{P}_1$  finite element method. We assume that all the angles of the triangles  $K_i \in \mathcal{T}_h$  are less than or equal to  $\pi/2$ . Show that  $u_h(x) \geq 0$  in  $\Omega$  if  $f(x) \geq 0$  in  $\Omega$ . Hint: we shall show that, for all  $\epsilon > 0$ ,  $\mathcal{K}_h + \epsilon \mathbf{I}$  is an M-matrix, where  $\mathcal{K}_h$  is the stiffness matrix.

There is obviously no difficulty to extend the  $\mathbb{P}_k$  finite element method to problems other than (6.36).

**Exercise 6.3.14** Apply the  $\mathbb{P}_k$  finite element method to the elasticity system (5.56). Show in particular that the stiffness matrix  $\mathcal{K}_h$  is in this case of order  $Nn_{dl}$  where  $N$  is the space dimensions and  $n_{dl}$  is the number of vertex degrees of freedom.

**Exercise 6.3.15** Make explicit the stiffness matrix  $\mathcal{K}_h$  obtained by application of the  $\mathbb{P}_k$  finite element method to the Neumann problem

$$\begin{cases} -\Delta u + au = f & \text{in } \Omega \\ \frac{\partial u}{\partial n} = g & \text{on } \partial\Omega, \end{cases} \quad (6.47)$$

with  $f \in L^2(\Omega)$ ,  $g \in L^2(\partial\Omega)$ , and  $a \in L^\infty(\Omega)$  such that  $a(x) \geq a_0 > 0$  a.e. in  $\Omega$ .

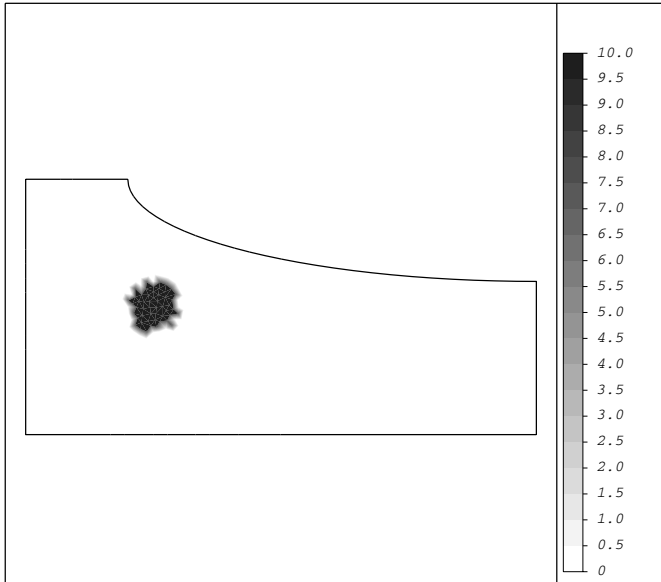


Figure 6.14. Source term  $f$  in the equation (6.36).

**Exercise 6.3.16** Show that the stiffness matrix  $\mathcal{K}_h$  obtained by application of the  $\mathbb{P}_k$  finite element method to the convection–diffusion problem of exercise 5.2.2 is invertible but not symmetric.

**Exercise 6.3.17** We propose to solve numerically the plate equation (5.70) by a (Hermite) finite element method in  $N = 2$  dimensions. For a triangular mesh  $\mathcal{T}_h$  we introduce the discrete space

$$V_h = \{v \in C^1(\overline{\Omega}) \text{ such that } v|_{K_i} \in \mathbb{P}_5 \text{ for all } K_i \in \mathcal{T}_h\}.$$

Show that every polynomial  $p \in \mathbb{P}_5$  is characterized uniquely on a triangle  $K$  by the following 21 real values

$$p(a_j), \quad \nabla p(a_j), \quad \nabla \nabla p(a_j), \quad \frac{\partial p(b_j)}{\partial n} \quad j = 1, 2, 3, \quad (6.48)$$

where  $(a_1, a_2, a_3)$  are the vertices of  $K$ ,  $(b_1, b_2, b_3)$  the middle of the sides of  $K$ , and  $\partial p(b_j)/\partial n$  denotes the derivative which is normal to the side of  $b_j$ . Show that  $V_h$  is a subspace of  $H^2(\Omega)$  whose elements  $v$  are uniquely characterized by the values (6.48) for each vertex and edge midpoint of the mesh. Deduce from this a finite element method (the Argyris method) to solve (5.70).

We finish this section by illustrating it with a numerical result obtained by the  $\mathbb{P}_1$  finite element method applied to the Dirichlet problem (6.36). The right-hand

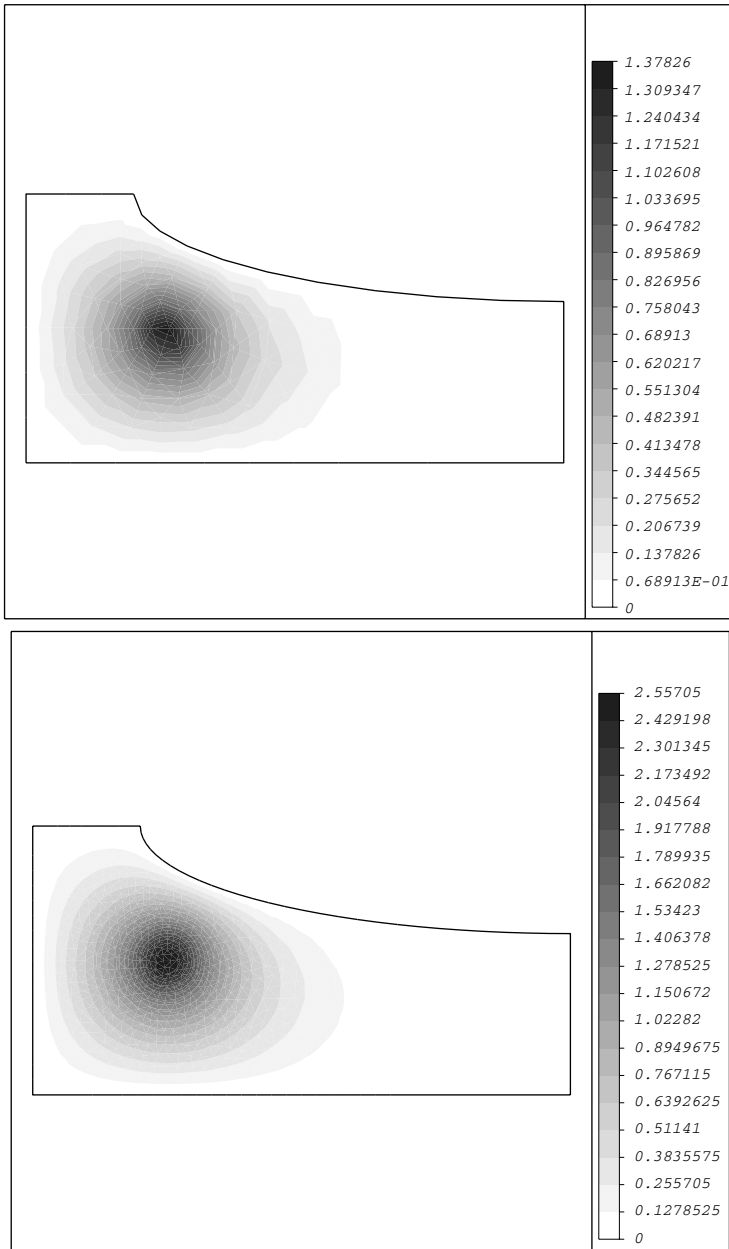


Figure 6.15. Approximate solution  $u_h$  of the diffusion equation (6.36) for the coarse mesh of figure 6.7 (top) and for the fine mesh of figure 6.13 (bottom).



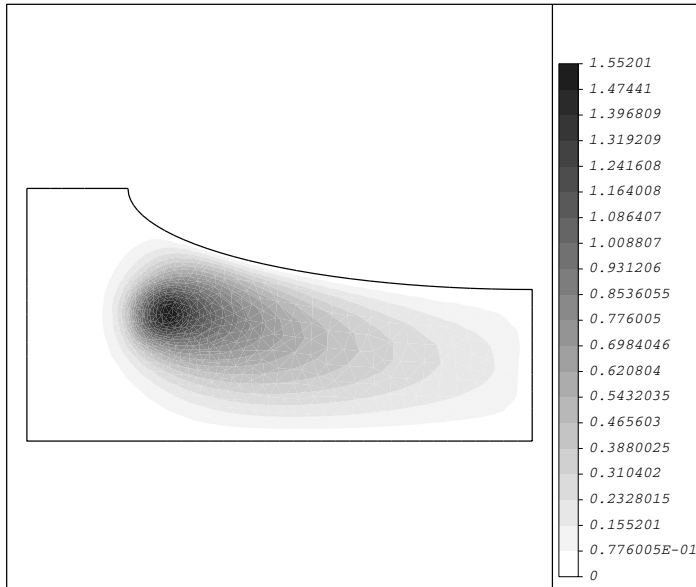


Figure 6.16. Approximate solution  $u_h$  of the convection–diffusion equation (5.13).

side is given by Figure 6.14. We can interpret this problem as the modelling of the diffusion in the atmosphere of a pollutant emitted by a localized source. The domain of calculation represents a region around the source (the vertical direction is ‘averaged’ and absent from the calculation) and we assume that the concentration is zero on the boundary. We have used the ‘coarse’ mesh of Figure 6.7, and the ‘fine’ mesh of Figure 6.13. The corresponding results are shown in Figure 6.15. We remark that the maximum value of the numerical solution  $u_h$  is higher for the fine mesh than for the coarse mesh (the scales are not the same). This is a manifestation of the fact that the step  $h$  of the coarse mesh is not small enough for the approximate solution  $u_h$  to have converged to the exact solution. If we add, as well as the diffusion, a convection effect (modelling a constant wind in the horizontal direction, see (5.13)), we can see the plume effect produced on the concentration in Figure 6.16 (obtained with the fine mesh). The maximal value of the solution is smaller in the presence of a convection term, this correspond to the physical intuition that the wind ‘dilutes’ the higher concentrations of pollutant.

### 6.3.2 Convergence and error estimation

We prove the convergence of  $\mathbb{P}_k$  finite element methods for the Dirichlet problem (6.36). We emphasize that this is only for a model problem, and that these methods converge for other problems, such as Neumann (6.47) problems. We shall need some geometric hypotheses on the quality of the mesh. For every  $N$ -simplex  $K$  we introduce

two geometric parameters: the **diameter**  $\text{diam}(K)$  and the **diameter of the largest ball contained in  $K$**   $\rho(K)$  (see Figure 6.17),

$$\text{diam}(K) = \max_{x,y \in K} \|x - y\|, \quad \rho(K) = \max_{B_r \subset K} (2r).$$

Of course, we always have  $\text{diam}(K)/\rho(K) > 1$ . This ratio is larger when  $K$  is ‘flat-tened’: it is a measure of the degeneracy of  $K$ . In practice, as in theory, we must avoid the use of  $N$ -simplices  $K$  which are too flat.

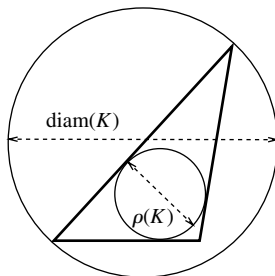


Figure 6.17.  $\text{diam}(K)$  and  $\rho(K)$  for a triangle  $K$ .

**Definition 6.3.11** Let  $(\mathcal{T}_h)_{h>0}$  be a sequence of meshes of  $\Omega$ . We say that it is a sequence of **regular meshes** if

1. the sequence  $h = \max_{K_i \in \mathcal{T}_h} \text{diam}(K_i)$  tends to 0,
2. there exists a constant  $C$  such that, for all  $h > 0$  and all  $K \in \mathcal{T}_h$ ,

$$\frac{\text{diam}(K)}{\rho(K)} \leq C. \quad (6.49)$$

**Remark 6.3.12** In  $N = 2$  dimensions the condition (6.49) is equivalent to the following condition on the angles of the triangle  $K$ : there exists a minimum angle  $\theta_0 > 0$  which is a lower bound (uniformly in  $h$ ) for all the angles of every  $K \in \mathcal{T}_h$ . We emphasize the fact that condition (6.49) is as important in practice as for the convergence analysis which follows. •

We can now state the principal result of this section which states the convergence of the  $\mathbb{P}_k$  finite element method and which gives an estimate of the rate of convergence if the solution is regular.

**Theorem 6.3.13** Let  $(\mathcal{T}_h)_{h>0}$  be a sequence of regular meshes of  $\Omega$ . Let  $u \in H_0^1(\Omega)$  be the solution of the Dirichlet problem (6.36), and  $u_h \in V_{0h}$ , its internal approximation

(6.42) by the  $\mathbb{P}_k$  finite element method. Then the  $\mathbb{P}_k$  finite element method converges, that is,

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(\Omega)} = 0. \quad (6.50)$$

Moreover, if  $u \in H^{k+1}(\Omega)$  and if  $k+1 > N/2$ , then we have the error estimate

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^k \|u\|_{H^{k+1}(\Omega)}, \quad (6.51)$$

where  $C$  is a constant independent of  $h$  and of  $u$ .

**Remark 6.3.14** Theorem 6.3.13 in fact applies to all Lagrange finite element methods (for example, the rectangular finite elements of Section 6.3.3). In effect, the only argument used is the construction of an interpolation operator based on the characterization of the functions of  $V_h$  by their values at the vertices of the degrees of freedom, which is always possible for Lagrange finite elements (see remark 6.3.6). Let us remark that, for the physically pertinent cases  $N = 2$  or  $N = 3$ , the condition  $k+1 > N/2$  is always satisfied as  $k \geq 1$ . •

**Remark 6.3.15** The error estimate (6.51) of theorem 6.3.13 is only true if the exact solution  $u$  is regular, which is not always the case. If  $u$  is not regular, we find in practice that the convergence is slower (see Figure 6.5 in one space dimension). On the other hand, the convergence (6.50), which is in the ‘energy’ space, does not imply the point convergence of  $u_h$  or of its derivatives. Figure 6.18 illustrates this fact for the Dirichlet problem (6.36) with  $f \equiv 1$  and a ‘reentrant corner’ where the solution is singular (see lemma 5.2.33). Numerically, the modulus of the gradient of  $u_h$  increases to infinity in the corner as  $h$  tends to zero (the maximum of  $|\nabla u_h|$  becomes 0.92 for the mesh on the left with 1187 vertices, 1.18 for the mesh in the middle with 4606 vertices, and 1.50 for that on the right with 18572 vertices) •

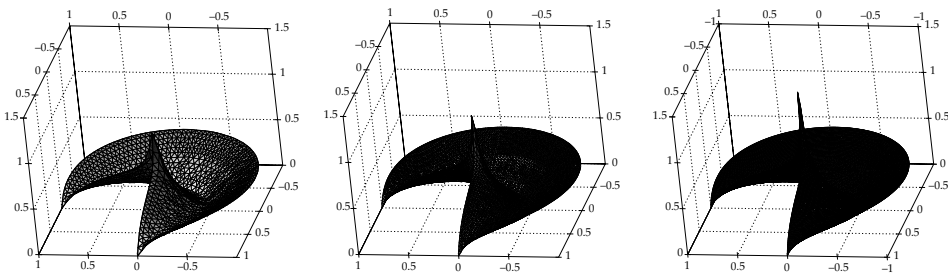


Figure 6.18. Modulus of the gradient of  $u_h$  for three meshes (increasingly fine from left to right).

The proof of theorem 6.3.13 rests on the following definition of an **interpolation operator**  $r_h$  and on the interpolation result of proposition 6.3.16. Let us recall that

we have denoted by  $(\hat{a}_i)_{1 \leq i \leq n_{dl}}$  the family of the vertices of the degrees of freedom and  $(\phi_i)_{1 \leq i \leq n_{dl}}$  the basis of  $V_{0h}$  of the  $\mathbb{P}_k$  finite element method (see proposition 6.3.7). For every continuous function  $v$ , we define its interpolant

$$r_h v(x) = \sum_{i=1}^{n_{dl}} v(\hat{a}_i) \phi_i(x). \quad (6.52)$$

The principal difference with the study made in  $N = 1$  dimension is that as the functions of  $H^1(\Omega)$  are not continuous when  $N \geq 2$ , the interpolation operator  $r_h$  is not defined on  $H^1(\Omega)$  (the point values of a function of  $H^1(\Omega)$  do not *a priori* have a meaning). Nevertheless, and this is the reason for the hypothesis  $k + 1 > N/2$ ,  $r_h$  is well defined on  $H^{k+1}(\Omega)$  because the functions of  $H^{k+1}(\Omega)$  are continuous ( $H^{k+1}(\Omega) \subset C(\overline{\Omega})$  from theorem 4.3.25).

**Proposition 6.3.16** *Let  $(\mathcal{T}_h)_{h>0}$  be a sequence of regular meshes of  $\Omega$ . We assume that  $k + 1 > N/2$ . Then, for all  $v \in H^{k+1}(\Omega)$  the interpolant  $r_h v$  is well defined, and there exists a constant  $C$ , independent of  $h$  and of  $v$ , such that*

$$\|v - r_h v\|_{H^1(\Omega)} \leq C h^k \|v\|_{H^{k+1}(\Omega)}. \quad (6.53)$$

Assuming, for the moment, proposition 6.3.16, we can conclude as for the convergence of the  $\mathbb{P}_k$  finite element method.

**Proof of theorem 6.3.13.** We apply the abstract framework of section 6.1.2. To show (6.50) we use lemma 6.1.3 with  $\mathcal{V} = C_c^\infty(\Omega)$  which is dense in  $H_0^1(\Omega)$ . As  $C_c^\infty(\Omega) \subset H^{k+1}(\Omega)$ , the estimate (6.53) of proposition 6.3.16 allows us to verify the hypothesis (6.5) of lemma 6.1.3 (for regular functions we do not need the condition  $k + 1 > N/2$  in the Proposition 6.3.16).

To obtain the error estimate (6.51) we use C  a's lemma 6.1.2 which told us that

$$\|u - u_h\|_{H^1(\Omega)} \leq C \inf_{v_h \in V_{0h}} \|u - v_h\|_{H^1(\Omega)} \leq C \|u - r_h u\|_{H^1(\Omega)},$$

if  $r_h u$  belongs to  $H^1(\Omega)$ . By application of proposition 6.3.16 to  $u$  we obtain (6.51).  $\square$

**Remark 6.3.17** The theorem 6.3.13 is valid when  $u_h$  is the **exact** solution of the internal approximation (6.42) in  $V_{0h}$ . This means exactly calculating all the integrals occurring in the matrix  $\mathcal{K}_h$  and the right-hand side  $b_h$ . In practice, we do not evaluate them exactly because we use numerical integration. Nevertheless, if we use ‘reasonable’ quadrature formulas, the  $\mathbb{P}_k$  finite element method converges (see [34]). In particular, if the quadrature formula used to calculate the integrals on an  $N$ -simplex  $K$  is exact for polynomials of  $\mathbb{P}_{2k-2}$ , then the error estimate (6.51) is always valid (where  $u_h$  is the discrete solution calculated with numerical integration). For example, for  $\mathbb{P}_1$  finite elements we can use the quadrature formulas (6.44) or (6.45) without loss of accuracy or rate of convergence.  $\bullet$

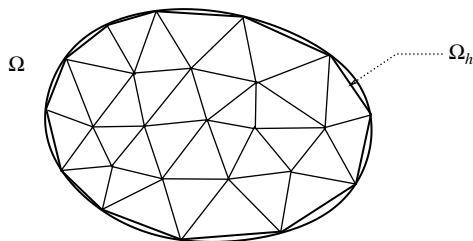


Figure 6.19. Approximation by a polyhedral domain  $\Omega_h$  of a regular open set  $\Omega$ .

**Remark 6.3.18** Let us indicate briefly what happens when the domain  $\Omega$  is not polyhedral (but sufficiently regular). We start by approximating  $\Omega$  by a polyhedral domain  $\Omega_h$  which we mesh by  $\mathcal{T}_h$  (see Figure 6.19). We can choose  $\Omega_h$  and its mesh (with  $h$  the maximum diameter of the elements) in such a way that there exists a constant  $C$  (which only depends on the curvature of  $\Omega$ ) satisfying

$$\text{dist}(\partial\Omega, \partial\Omega_h) \leq Ch^2.$$

We call  $u_h$  the solution of the variational approximation in the space  $V_h$  associated with the mesh  $\mathcal{T}_h$  and the  $\mathbb{P}_k$  finite element. In general, even if we choose  $\Omega_h \subset \Omega$ ,  $V_h$  is **not a subspace** of the Sobolev space  $V$  in which we look for the exact solution  $u$  (for example, if the boundary conditions are Neumann), which seriously complicates the analysis. Nevertheless, in  $N = 2$  dimensions we can show (see [34]) that, for  $\mathbb{P}_1$  finite elements, if  $u \in H^2(\Omega)$ , then we always have

$$\|u - u_h\|_{H^1(\Omega_h)} \leq Ch\|u\|_{H^2(\Omega)}, \quad (6.54)$$

while, for  $\mathbb{P}_2$  finite elements, if  $u \in H^{k+1}(\Omega)$ , then we only have

$$\|u - u_h\|_{H^1(\Omega_h)} \leq Ch^{3/2}\|u\|_{H^{k+1}(\Omega)}. \quad (6.55)$$

Consequently, this method is satisfactory for  $\mathbb{P}_1$  finite elements, since the convergence (6.54) is of the same order as (6.51), but disappointing and nonoptimal for  $\mathbb{P}_k$  finite elements with  $k \geq 2$ . We can fix this situation by introducing ‘isoparametric finite elements’: they work by meshing the part of  $\Omega$  near to the boundary by ‘curves’ obtained by deformation of standard  $N$ -simplices (this deformation is a generalization of the affine transformation introduced in remark 6.3.10). For example, in  $N = 2$  dimensions we often use a polynomial transformation of degree 2 which deforms a reference triangle into a ‘triangle’ whose sides are arcs of a parabola. This allows a better approximation of the boundary  $\partial\Omega$  by  $\partial\Omega_h$ . We can then prove an optimal error estimate of the same order as (6.51) (see [9], [34]). •

We pass now to the proof of proposition 6.3.16 which can be omitted in first reading. It works by the construction of a local interpolation operator in each element of the mesh. Let  $K$  be an  $N$ -simplex of lattice of order  $k$ ,  $\Sigma_k$ . We define the interpolation operator  $r_K$ , for every continuous function  $v$  on  $K$ ,

$$r_K v = p \in \mathbb{P}_k \text{ such that } p(x) = v(x) \quad \forall x \in \Sigma_k. \quad (6.56)$$

From lemma 6.3.3 we know that every polynomial of  $\mathbb{P}_k$  is uniquely determined by its values at the points of  $\Sigma_k$ : consequently, (6.56) defines  $r_K$  as a (linear) mapping.

**Lemma 6.3.19 (Bramble–Hilbert)** *We assume that  $k + 1 > N/2$ . The interpolation operator  $r_K$  is linear and continuous from  $H^{k+1}(K)$  into  $H^{k+1}(K)$ , and there exists a constant  $C(K)$  such that, for all  $v \in H^{k+1}(K)$  we have*

$$\|v - r_K v\|_{H^{k+1}(K)} \leq C(K) |v|_{H^{k+1}(K)}, \quad (6.57)$$

where  $|v|_{H^{k+1}(K)}$  is the semi-norm defined by

$$|v|_{H^{k+1}(K)}^2 = \sum_{|\alpha|=k+1} \int_K |\partial^\alpha v|^2 dx = \|v\|_{H^{k+1}(K)}^2 - \|v\|_{H^k(K)}^2.$$

**Proof.** For  $k + 1 > N/2$  theorem 4.3.25 means that  $H^{k+1}(K) \subset C(K)$ , therefore the point values of the functions of  $H^{k+1}(K)$  are well defined as continuous linear forms. Consequently,  $r_K v$  is a polynomial whose coefficients depend linearly and continuously on  $v \in H^{k+1}(K)$ , in any space  $H^m(K)$  with  $m \in \mathbb{N}$ . We deduce that  $r_K$  is linear and continuous in  $H^{k+1}(K)$ . Let us now prove the inequality

$$\|v\|_{H^{k+1}(K)} \leq C(K) (|v|_{H^{k+1}(K)} + \|r_K v\|_{H^{k+1}(K)}), \quad (6.58)$$

proceeding by contradiction (as we have already done for other inequalities; see, for example, the proof (4.15) of the Poincaré inequality). There therefore exists a sequence  $v_n \in H^{k+1}(K)$  such that

$$1 = \|v_n\|_{H^{k+1}(K)} > n (|v_n|_{H^{k+1}(K)} + \|r_K v_n\|_{H^{k+1}(K)}). \quad (6.59)$$

The left-hand term of (6.59) implies that the sequence  $v_n$  is bounded in  $H^{k+1}(K)$ . By application of Rellich's theorem 4.3.21, there exists a subsequence  $v_{n'}$  which converges in  $H^k(K)$ . The right-hand term of (6.59) implies that the sequence of derivatives  $\partial^\alpha v_{n'}$ , for every multi-index  $|\alpha| = k + 1$ , converges to zero in  $L^2(K)$ . Consequently,  $v_{n'}$  converges in  $H^{k+1}(K)$  to a limit  $v$  which satisfies (by passing to the limit in (6.59))

$$|v|_{H^{k+1}(K)} = 0, \quad \|r_K v\|_{H^{k+1}(K)} = 0. \quad (6.60)$$

The first equation of (6.60) shows that  $v \in \mathbb{P}_k$  because  $K$  is connected (by repeated application of proposition 4.2.5). By the definition (6.56) of  $r_K$  we have  $r_K v = v$  for  $v \in \mathbb{P}_k$ . The second equation of (6.60) shows therefore that  $r_K v = v = 0$ , which is a contradiction with the limit of the left-hand term of (6.59). To obtain (6.57) we apply (6.58) to  $(v - r_K v)$  by remarking that  $r_K(v - r_K v) = 0$  and that  $|v - r_K v|_{H^{k+1}(K)} = |v|_{H^{k+1}(K)}$  since the derivatives of order  $k + 1$  of a polynomial of  $\mathbb{P}_k$  are zero.  $\square$

The disadvantage of the Bramble–Hilbert lemma 6.3.19 is that the constant in the inequality (6.57) depends on  $K$  in a nonexplicit way. We make this dependence precise in the following lemma.

**Lemma 6.3.20** *We assume that  $k + 1 > N/2$  and that  $\text{diam}(K) \leq 1$ . There exists a constant  $C$  independent of  $K$  such that, for all  $v \in H^{k+1}(K)$  we have*

$$\|v - r_K v\|_{H^1(K)} \leq C \frac{(\text{diam}(K))^{k+1}}{\rho(K)} |v|_{H^{k+1}(K)}. \quad (6.61)$$

**Proof.** We use remark 6.3.10 which confirms that every  $N$ -simplex  $K$  is the image by an affine transformation of the reference  $N$ -simplex  $K_0$ , defined by (6.46). In other words, there exists an invertible matrix  $B$  and a vector  $b$  (depending on  $K$ ) such that, for all  $x \in K$ , there exists  $x_0 \in K_0$  satisfying

$$x = Bx_0 + b. \quad (6.62)$$

To obtain (6.61), we start from the inequality (6.57) established on  $K_0$  and we apply the change of variable (6.62). This allows us to find the dependence with respect to  $K$  of the constant in this inequality. We shall not give the details of this calculation but simply indicate the principal stages (the reader can consult [34]). The Jacobian of the change of variable being  $\det(B)$ , and the derivatives in  $K$  being obtained from the derivatives in  $K_0$  by composition with  $B^{-1}$ , there exists a constant  $C$ , independent of  $K$ , such that, for every regular function  $v(x)$  with  $v_0(x_0) = v(Bx_0 + b)$ , we have

$$\begin{aligned} |v_0|_{H^l(K_0)} &\leq C \|B\|^l |\det(B)|^{-1/2} |v|_{H^l(K)} \\ |v|_{H^l(K)} &\leq C \|B^{-1}\|^l |\det(B)|^{1/2} |v_0|_{H^l(K_0)}. \end{aligned}$$

We therefore deduce from (6.57)

$$\begin{aligned} |v - r_K v|_{H^1(K)} &\leq C \|B\|^{k+1} \|B^{-1}\| |v|_{H^{k+1}(K)} \\ \|v - r_K v\|_{L^2(K)} &\leq C \|B\|^{k+1} |v|_{H^{k+1}(K)}. \end{aligned}$$

In addition, we easily verify that

$$\|B\| \leq \frac{\text{diam}(K)}{\rho(K_0)}, \quad \|B^{-1}\| \leq \frac{\text{diam}(K_0)}{\rho(K)}.$$

Combining these results we obtain (6.61).  $\square$

**Proof of proposition 6.3.16.** By construction, if  $v \in H^{k+1}(\Omega)$ , its interpolant  $r_h v$  restricted to the  $N$ -simplex  $K$  is simply  $r_K v$ . Consequently,

$$\|v - r_h v\|_{H^1(\Omega)}^2 = \sum_{K_i \in \mathcal{T}_h} \|v - r_{K_i} v\|_{H^1(K_i)}^2.$$

We apply the upper bound (6.61) to each element  $K_i$  (with the same constant  $C$  for all), and as the mesh is regular the inequality (6.49) allows us to uniformly bound the ratio  $\text{diam}(K_i)/\rho(K_i)$ . We deduce

$$\|v - r_h v\|_{H^1(\Omega)}^2 \leq Ch^{2k} \sum_{K_i \in \mathcal{T}_h} |v|_{H^{k+1}(K_i)}^2 \leq Ch^{2k} \|v\|_{H^{k+1}(\Omega)}^2$$

which is the desired result.  $\square$

**Exercise 6.3.18** Show that for a sequence of regular meshes, and for  $\mathbb{P}_1$  finite elements, the interpolation operator  $r_h$  satisfies in  $N = 2$  or 3 dimensions

$$\|v - r_h v\|_{L^2(\Omega)} \leq Ch^2 \|v\|_{H^2(\Omega)}.$$

### 6.3.3 Rectangular finite elements

If the domain  $\Omega$  is rectangular (that is,  $\Omega$  is a polyhedral open set whose faces are perpendicular to the axes), we can mesh by rectangles (see Figure 6.20) and use an adapted finite element method. We shall define the Lagrange finite elements (that is, whose degrees of freedom are point values of functions), called  $\mathbb{Q}_k$  finite elements. Let us start by defining an  $N$ -rectangle  $K$  of  $\mathbb{R}^N$  as the (nondegenerate) block  $\prod_{i=1}^N [l_i, L_i]$  with  $-\infty < l_i < L_i < +\infty$ . We denote by  $(a_j)_{1 \leq j \leq 2^N}$  the vertices of  $K$ .

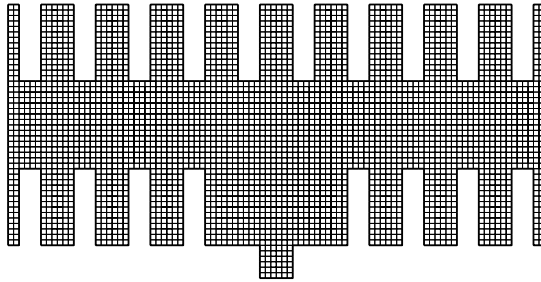


Figure 6.20. Example of a rectangular mesh in  $N = 2$  dimensions.

**Definition 6.3.21** Let  $\Omega$  be a polyhedral connected open set of  $\mathbb{R}^N$ . A **rectangular mesh** of  $\overline{\Omega}$  is a set  $\mathcal{T}_h$  of (nondegenerate)  $N$ -rectangles  $(K_i)_{1 \leq i \leq n}$  which satisfies

1.  $K_i \subset \overline{\Omega}$  and  $\overline{\Omega} = \bigcup_{i=1}^n K_i$ ,
2. the intersection  $K_i \cap K_j$  of two distinct  $N$ -rectangles is an  $m$ -rectangle, with  $0 \leq m \leq N - 1$ , whose vertices are also vertices of  $K_i$  and  $K_j$ . (In  $N = 2$  dimensions, the intersection of two rectangles is either empty, or a common vertex, or an entire common face.)

The **vertices** or **nodes** of the mesh  $\mathcal{T}_h$  are the vertices of the  $N$ -rectangles  $K_i$ . By convention, the parameter  $h$  denotes the maximum diameter of the  $N$ -rectangles  $K_i$ .

We define the set  $\mathbb{Q}_k$  of polynomials with real coefficients from  $\mathbb{R}^N$  into  $\mathbb{R}$  of degree less than or equal to  $k$  **with respect to each variable**, that is, for all  $p \in \mathbb{Q}_k$  written in the form

$$p(x) = \sum_{0 \leq i_1 \leq k, \dots, 0 \leq i_N \leq k} \alpha_{i_1, \dots, i_N} x_1^{i_1} \cdots x_N^{i_N} \quad \text{with } x = (x_1, \dots, x_N).$$

Let us remark that the total degree of  $p$  can be greater than  $k$ , which differentiates the space  $\mathbb{Q}_k$  from  $\mathbb{P}_k$ .



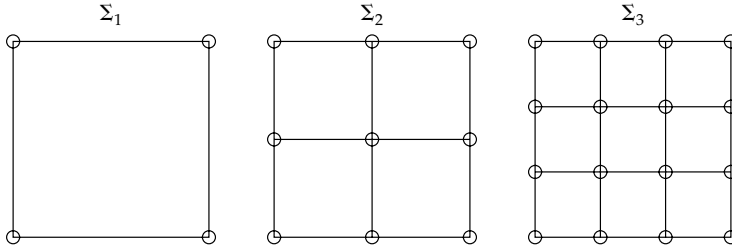


Figure 6.21. Lattice of order 1, 2, and 3 for a rectangle (the circles represent the points of the lattice).

For every integer  $k \geq 1$  we define the **lattice of order  $k$**  of the  $N$ -rectangle  $K$  as the set

$$\Sigma_k = \left\{ x \in K \text{ such that } \frac{x_j - l_j}{L_j - l_j} \in \left\{ 0, \frac{1}{k}, \dots, \frac{k-1}{k}, 1 \right\} \text{ for } 1 \leq j \leq N \right\}. \quad (6.63)$$

For  $k = 1$  this is the set of the vertices of  $K$ , and for  $k = 2$  and  $N = 2$  of the vertices, of the midpoints of the edges linking two vertices, and of the barycentre (see Figure 6.21).

The lattice  $\Sigma_k$  of an  $N$ -rectangle  $K$  is **unisolvant** for  $\mathbb{Q}_k$ , that is, we can characterize all the polynomials of  $\mathbb{Q}_k$ .

**Lemma 6.3.22** *Let  $K$  be an  $N$ -rectangle. Take an integer  $k \geq 1$ . Then, every polynomial of  $\mathbb{Q}_k$  is uniquely determined by its values at the points of the lattice of order  $k$ ,  $\Sigma_k$ , defined by (6.63).*

**Proof.** We verify that the cardinality of  $\Sigma_k$  and the dimension of  $\mathbb{Q}_k$  coincide

$$\text{card}(\Sigma_k) = \dim(\mathbb{Q}_k) = (k+1)^N.$$

Since the mapping which, for every polynomial of  $\mathbb{Q}_k$ , gives its values on the lattice  $\Sigma_k$  is linear, it is sufficient to show a basis of  $\mathbb{Q}_k$  whose elements are 1 at a point of the lattice and 0 elsewhere to prove the result. Let  $x^\mu$  be a point of  $\Sigma_k$  defined by

$$\frac{x_j^\mu - l_j}{L_j - l_j} = \frac{\mu_j}{k} \quad \text{with } 0 \leq \mu_j \leq k, \quad \forall j \in \{1, \dots, N\}.$$

We define the polynomial  $p \in \mathbb{Q}_k$  by

$$p(x) = \prod_{j=1}^N \left( \prod_{\substack{i=0 \\ i \neq \mu_j}}^k \frac{k(x_j - l_j) - i(L_j - l_j)}{(\mu_j - i)(L_j - l_j)} \right) \quad \text{with } x = (x_1, \dots, x_N).$$

We verify easily that  $p(x^\mu) = 1$  while  $p$  is zero on all the other points of  $\Sigma_k$ , which is the desired result.  $\square$

As in the triangular case we have the following continuity condition across a face (we leave the proof, similar to that of lemma 6.3.4, to the reader).

**Lemma 6.3.23** *Let  $K$  and  $K'$  be two  $N$ -rectangles having a common face  $\Gamma = \partial K \cap \partial K'$ . Let  $k \geq 1$  be an integer. Then, their lattice of order  $k$   $\Sigma_k$  and  $\Sigma'_k$  coincide on this face  $\Gamma$ . Moreover, given  $p_K$  and  $p_{K'}$ , two polynomials of  $\mathbb{Q}_k$ , the function  $v$  is defined by*

$$v(x) = \begin{cases} p_K(x) & \text{if } x \in K \\ p_{K'}(x) & \text{if } x \in K' \end{cases}$$

*is continuous on  $K \cup K'$ , if and only if  $p_K$  and  $p_{K'}$  have values which coincide at the points of the lattice on the common face  $\Gamma$ .*

In practice, we mostly use the spaces  $\mathbb{Q}_1$  and  $\mathbb{Q}_2$ . Figure 6.22 shows a function of  $\mathbb{Q}_1$  in  $N = 2$  dimensions (we can verify that the functions of  $\mathbb{Q}_1$  are not piecewise affine like those of  $\mathbb{P}_1$ ).

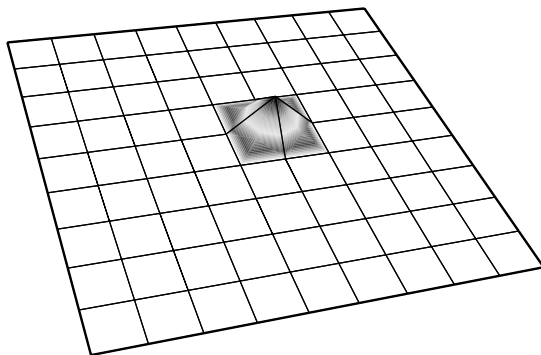


Figure 6.22. Function of  $\mathbb{Q}_1$  in  $N = 2$  dimensions.

**Exercise 6.3.19** Let  $K = [0, 1]^2$  be the unit cube in  $N = 2$  dimensions with vertices  $a^1 = (0, 0)$ ,  $a^2 = (1, 0)$ ,  $a^3 = (1, 1)$ ,  $a^4 = (0, 1)$ . We define  $x_3 = 1 - x_1$ ,  $x_4 = 1 - x_2$ , and  $\bar{i}$  as the value of  $i$  modulo 4. Verify that the basis functions of  $\mathbb{Q}_1$  are

$$p_i(x) = x_{\bar{i}+2} x_{\bar{i}+3} \quad \text{for } 1 \leq i \leq 4,$$

and that those of  $\mathbb{Q}_2$  are

$$\begin{aligned} P_i(x) &= x_{\bar{i}+2}(2x_{\bar{i}+2} - 1)x_{\bar{i}+3}(2x_{\bar{i}+3} - 1) & \text{for } 1 \leq i \leq 4 \\ P_i(x) &= -4x_{\bar{i}+2}(x_{\bar{i}+2} - 1)x_{\bar{i}+3}(2x_{\bar{i}+3} - 1) & \text{for } 5 \leq i \leq 8 \\ P_9(x) &= 16x_1x_2x_3x_4. \end{aligned}$$

**Remark 6.3.24** In practice, we sometimes replace the  $\mathbb{Q}_2$  finite element by another finite element which is more simple, and just as effective, denoted by  $\mathbb{Q}_2^*$ . In  $N = 2$  dimensions, the  $\mathbb{Q}_2^*$  finite element is defined by the 8 basis functions  $(p_i)_{1 \leq i \leq 8}$  of exercise 6.3.19 (we have removed the last,  $p_9$ ). We verify that the degrees of freedom of  $\mathbb{Q}_2^*$  are the vertices and the middle of the edges of the rectangle (but not its barycentre). In  $N = 3$  dimensions, the  $\mathbb{Q}_2^*$  finite element is defined by its degrees of freedom which are the 8 vertices and the 12 middles of the edges of the cube (there are no degrees of freedom in the interior). •

**Definition 6.3.25** Given a rectangular mesh  $\mathcal{T}_h$  of an open set  $\Omega$ , the  $\mathbb{Q}_k$  finite element method is defined by the discrete space

$$V_h = \{v \in C(\bar{\Omega}) \text{ such that } v|_{K_i} \in \mathbb{Q}_k \text{ for all } K_i \in \mathcal{T}_h\}. \quad (6.64)$$

The nodes of these **degrees of freedom** are the set of the points  $(\hat{a}_i)_{1 \leq i \leq n_{dl}}$  of the lattice of order  $k$  of each of the  $N$ -rectangles  $K_i \in \mathcal{T}_h$ .

As in the triangular case, definition 6.3.25 has a meaning, thanks to the following proposition (whose proof we shall leave to the reader as an exercise).

**Proposition 6.3.26** The space  $V_h$ , defined by (6.64), is a subspace of  $H^1(\Omega)$  whose dimension is the number of degrees of freedom  $n_{dl}$ . Moreover, there exists a basis of  $V_h$   $(\phi_i)_{1 \leq i \leq n_{dl}}$  defined by

$$\phi_i(\hat{a}_j) = \delta_{ij} \quad 1 \leq i, j \leq n_{dl},$$

such that

$$v(x) = \sum_{i=1}^{n_{dl}} v(\hat{a}_i) \phi_i(x).$$

As the  $\mathbb{Q}_k$  finite elements are Lagrange finite elements, we can prove the same convergence results as for the method of  $\mathbb{P}_k$  finite elements. We shall allow the reader to verify that the proof of theorem 6.3.13 applies ‘mutatis mutandis’ to the following theorem (the definition 6.3.11 of regular meshes is easily extended to rectangular meshes).

**Theorem 6.3.27** Let  $(\mathcal{T}_h)_{h>0}$  be a sequence of regular rectangular meshes of  $\Omega$ . Let  $u \in H_0^1(\Omega)$  be the exact solution of the Dirichlet problem (6.36), and  $u_h \in V_{0h}$ , the approximate solution by the  $\mathbb{Q}_k$  finite element method. Then the finite element method  $\mathbb{Q}_k$  converges, that is,

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(\Omega)} = 0.$$

Moreover, if  $u \in H^{k+1}(\Omega)$  and if  $k+1 > N/2$ , then we have the error estimate

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^k \|u\|_{H^{k+1}(\Omega)},$$

where  $C$  is a constant independent of  $h$  and of  $u$ .

**Remark 6.3.28** We can generalize the concept of a rectangular mesh and of  $\mathbb{Q}_k$  finite elements a little by using the concept of affine transformation. We say  $N$ -parallelotope to mean the image by an affine mapping  $F$  of the unit cube  $[0, 1]^N$  (a 2-parallelotope is a parallelogram). We verify that an  $N$ -parallelotope with  $2N$  faces, parallel in pairs, and that its lattice is the image of the lattice of the unit cube. We can then mesh a domain  $\Omega$  by  $N$ -parallelotopes and define a finite element method based on the image  $F(\mathbb{Q}_k)$ , in each  $N$ -parallelotope, of the space  $\mathbb{Q}_k$  for the unit cube (in general  $F(\mathbb{Q}_k) \neq \mathbb{Q}_k$ ). We can also use more complicated (nonaffine) transformations: this is the isoparametric finite element method (see the remark 6.3.18 on this subject). For example, in  $N = 2$  dimensions, the use of transformations  $\mathbb{Q}_1$  allows us to mesh a domain with arbitrary quadrilaterals (with nonparallel faces). For more details, we refer to [34]. •

**Remark 6.3.29** We can also mesh a part of  $\Omega$  by  $N$ -simplices, and another by  $N$ -rectangles and construct a finite element method which is a mixture of the two types  $\mathbb{P}_k$  and  $\mathbb{Q}_k$ . For more details, we again refer to [34]. •

**Remark 6.3.30** We can define finite elements intermediate between  $\mathbb{P}_k$  and  $\mathbb{Q}_k$  in  $N = 3$  dimensions, called ‘prismatic finite elements of order  $k$ ’. Assume that  $\Omega = \omega \times ]0, L[$  with  $\omega$  an open set of  $\mathbb{R}^2$ . We mesh  $\omega$  by triangles  $T_i$ , and  $]0, L[$  by segments  $[z_j, z_{j+1}]$ . We then define the prisms of  $\mathbb{R}^3$  as the product  $T_i \times [z_j, z_{j+1}]$ , with which we mesh  $\Omega$ . We then construct basis functions intermediate between those of  $\mathbb{P}_k$  and  $\mathbb{Q}_k$  on these prisms. For more details, we refer to [34]. •

### 6.3.4 Finite elements for the Stokes problem

The generalization of the finite element method to systems of partial differential equations (like the system of linear elasticity) does not pose particular problems. This is not the case for the system of Stokes equations (5.71) because of the incompressibility condition on the fluid (or zero divergence condition on the velocity). The considerable practical importance of numerical simulations in incompressible fluid mechanics justifies the fact that we shall briefly discuss this particular case (and also allows us to show that numerical analysis is not always the long tranquil river we might imagine from reading this text).

Recall that, in a connected bounded domain  $\Omega \subset \mathbb{R}^N$ , in the presence of exterior forces  $f(x)$ , and for boundary conditions which describe the adherence of the fluid to the boundary, the Stokes equations are written

$$\begin{cases} \nabla p - \mu \Delta u = f & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (6.65)$$

where  $\mu > 0$  is the viscosity of the fluid. In the Section 5.3.2 we have proposed as a variational formulation of (6.65)

$$\text{Find } u \in V \text{ such that } \int_{\Omega} \mu \nabla u \cdot \nabla v \, dx = \int_{\Omega} f \cdot v \, dx \quad \forall v \in V, \quad (6.66)$$

where  $V$  is the Hilbert space defined by

$$V = \{v \in H_0^1(\Omega)^N \text{ such that } \operatorname{div} v = 0 \text{ a.e. in } \Omega\}. \quad (6.67)$$

As  $V$  contains the incompressibility constraint  $\operatorname{div} v = 0$ , it is very difficult in practice to construct internal variational approximations of (6.66) as we have justified here. More precisely, the difficulty is to define simply (explicitly) a subspace  $V_h$  of  $V$  of finite dimension whose elements are written using the basis functions of  $\mathbb{P}_k$  or  $\mathbb{Q}_k$  finite elements. For example, if  $\mathcal{T}_h = (K_i)_{1 \leq i \leq n}$  is a triangular mesh of the open connected polyhedral set  $\Omega$ , we can define

$$V_h = \{v \in C(\overline{\Omega})^N \text{ such that } \operatorname{div} v = 0 \text{ in } \Omega, v|_{K_i} \in \mathbb{P}_k^N \text{ for all } K_i \in \mathcal{T}_h\},$$

but it is not clear that  $V_h$  is not ‘too small’ and how we can characterize its elements in terms of degrees of freedom. In particular, the condition  $\operatorname{div} v = 0$  in the definition of  $V_h$  mixes all the components of  $v$ , which makes it very difficult and complicated to characterize an explicit basis of  $V_h$ . We therefore do not use the variational formulation (6.67) to define a finite element method.

In practice, we introduce another variational formulation of the Stokes equations which consists of not forcing the incompressibility in the definition of the space and keeping the pressure as an unknown in the variational formulation. By multiplying the first equation of (6.65) by a test function  $v \in H_0^1(\Omega)^N$  and the second equation by another test function  $q \in L^2(\Omega)$ , we obtain after integration by parts: find  $(u, p) \in H_0^1(\Omega)^N \times L^2(\Omega)/\mathbb{R}$  such that

$$\begin{cases} \int_{\Omega} \mu \nabla u \cdot \nabla v \, dx - \int_{\Omega} p \operatorname{div} v \, dx = \int_{\Omega} f \cdot v \, dx \\ \int_{\Omega} q \operatorname{div} u \, dx = 0, \end{cases} \quad (6.68)$$

for all  $(v, q) \in H_0^1(\Omega)^N \times L^2(\Omega)/\mathbb{R}$ . A supplementary interest in (6.68) is that the pressure is not eliminated as in (6.66). It will therefore be possible to calculate this (physically important) variable with (6.68). We leave the following result to the reader to verify as an exercise.

**Lemma 6.3.31** *Take  $(u, p) \in H_0^1(\Omega)^N \times L^2(\Omega)/\mathbb{R}$ . The couple  $(u, p)$  is the solution of (6.68) if and only if it is the solution (weak, in the sense of theorem 5.3.8) of the Stokes equations (6.65).*

It is then easy to construct an internal variational approximation of (6.68). We introduce the discrete spaces

$$\begin{cases} V_{0h} = \{v \in C(\overline{\Omega})^N \text{ such that } v|_{K_i} \in \mathbb{P}_k^N \text{ for all } K_i \in \mathcal{T}_h \text{ and } v = 0 \text{ on } \partial\Omega\}, \\ Q_h = \{q \in C(\overline{\Omega})/\mathbb{R} \text{ such that } q|_{K_i} \in P_{k'} \text{ for all } K_i \in \mathcal{T}_h\}, \end{cases}$$

so that  $V_{0h} \times Q_h$  is a subspace of  $H_0^1(\Omega)^N \times L^2(\Omega)/\mathbb{R}$  of finite dimension. The internal variational approximation of (6.68) is simply

$$\begin{cases} \int_{\Omega} \mu \nabla u_h \cdot \nabla v_h \, dx - \int_{\Omega} p_h \operatorname{div} v_h \, dx = \int_{\Omega} f \cdot v_h \, dx \\ \int_{\Omega} q_h \operatorname{div} u_h \, dx = 0, \end{cases} \quad (6.69)$$

for all  $(v_h, q_h) \in V_{0h} \times Q_h$ . Let us explain how to solve (6.69) in practice. Denoting by  $n_V$  the dimension of  $V_{0h}$  and  $n_Q$  that of  $Q_h$ , we introduce the basis  $(\phi_j)_{1 \leq j \leq n_V}$  of  $V_{0h}$  and the basis  $(\psi_j)_{1 \leq j \leq n_Q}$  of  $Q_h$  constructed with the finite element basis functions (see Proposition 6.3.7). We decompose  $u_h$  and  $p_h$  in these bases

$$u_h(x) = \sum_{j=1}^{n_V} u_h(\hat{a}_j) \phi_j(x), \quad p_h(x) = \sum_{j=1}^{n_Q} u_h(\hat{a}'_j) \psi_j(x),$$

Denoting by  $U_h = (u_h(\hat{a}_j))_{1 \leq j \leq n_V}$  and  $P_h = (p_h(\hat{a}'_j))_{1 \leq j \leq n_Q}$ , we obtain the following linear system

$$\begin{pmatrix} A_h & B_h^* \\ B_h & 0 \end{pmatrix} \begin{pmatrix} U_h \\ P_h \end{pmatrix} = \begin{pmatrix} b_h \\ 0 \end{pmatrix}, \quad (6.70)$$

where  $B_h^*$  is the adjoint (or transposed) matrix of  $B_h$ ,  $b_h = (\int_{\Omega} f \cdot \phi_i \, dx)_{1 \leq i \leq n_V}$ , and

$$A_h = \left( \mu \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j \, dx \right)_{1 \leq i, j \leq n_V}, \quad B_h = \left( - \int_{\Omega} \psi_i \operatorname{div} \phi_j \, dx \right)_{1 \leq i \leq n_Q, 1 \leq j \leq n_V}.$$

Things become complicated when we need to know if we can always solve the linear system (6.70) uniquely. Let us remark that the matrix  $A_h$  is symmetric positive definite of order  $n_V$ , that the matrix  $B_h$  is rectangular of size  $n_Q \times n_V$ , and that, if the global matrix of (6.70) is symmetric of order  $n_V + n_Q$ , it is not positive definite. Nevertheless, we have the following result.

**Lemma 6.3.32** *The linear system (6.70) always has a solution  $(U_h, P_h)$  in  $\mathbb{R}^{n_V} \times \mathbb{R}^{n_Q}$ . The vector  $U_h$  is unique, while  $P_h$  is unique up to the addition of an element of  $\operatorname{Ker} B_h^*$ .*

**Proof.** Since  $(\operatorname{Ker} B_h)^\perp = \operatorname{Im} B_h^*$ , it is easy to see that (6.70) is equivalent to

$$\text{find } U_h \in \operatorname{Ker} B_h \text{ such that } A_h U_h \cdot W_h = b_h \cdot W_h \text{ for all } W_h \in \operatorname{Ker} B_h.$$

It is then sufficient to apply the Lax–Milgram theorem 3.3.1 to obtain the existence and uniqueness of  $U_h$  in  $\operatorname{Ker} B_h$ . Consequently, (6.70) has at least a solution  $(U_h, P_h)$  in  $\mathbb{R}^{n_V} \times \mathbb{R}^{n_Q}$ . As  $U_h$  must belong to  $\operatorname{Ker} B_h$ , it is unique in  $\mathbb{R}^{n_V}$ . In addition, we easily verify that  $P_h$  is unique up to the addition of an element of  $\operatorname{Ker} B_h^*$ .  $\square$

The complication is that the kernel  $\text{Ker} B_h^*$  is never reduced to the zero vector and it can sometimes be very ‘large’. Everything depends on the choice of the orders  $k$  and  $k'$  of the finite elements for the velocity and for the pressure.

**Lemma 6.3.33** *The kernel  $\text{Ker} B_h^*$  contains at least the vector  $\mathbb{I}$  of  $\mathbb{R}^{n_Q}$  all of whose components are equal to 1. In other words, the discrete pressure  $p_h$  is at best defined up to a constant.*

**Proof.** Take  $r_h \in Q_h$  and  $w_h \in V_{0h}$ . By definition

$$W_h \cdot B_h^* R_h = B_h W_h \cdot R_h = \int_{\Omega} r_h \text{div} w_h \, dx.$$

Now  $r_h = 1$  always belongs to  $Q_h$ , and since

$$\int_{\Omega} \text{div} w_h \, dx = \int_{\partial\Omega} w_h \cdot n \, ds = 0$$

for all  $w_h \in V_{0h}$ , we deduce that  $R_h = \mathbb{I} = (1, \dots, 1)$  belongs to  $\text{Ker} B_h^*$ .  $\square$

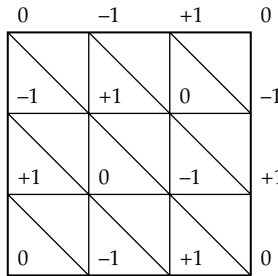


Figure 6.23. Unstable mode of the pressure on a uniform triangular mesh ( $\mathbb{P}_1$  finite elements for the velocity and the pressure).

**Lemma 6.3.34** *When  $k = 2$  and  $k' = 1$  ( $\mathbb{P}_2$  finite elements for the velocity and  $\mathbb{P}_1$  for the pressure), the kernel  $\text{Ker} B_h^*$  has dimension one, generated by the vector  $\mathbb{I}$  (in other words, the discrete pressure  $p_h$  is unique up to a constant).*

*When  $k = k' = 1$  ( $\mathbb{P}_1$  finite elements for the velocity and the pressure), the kernel  $\text{Ker} B_h^*$  is in general of dimension strictly larger than one (in other words, the discrete pressure  $p_h$  is not unique, even up to a constant).*

**Proof.** Take  $r_h \in Q_h$  and  $w_h \in V_{0h}$ . By definition

$$W_h \cdot B_h^* R_h = B_h W_h \cdot R_h = \int_{\Omega} r_h \text{div} w_h \, dx = - \int_{\Omega} \nabla r_h \cdot w_h \, dx.$$

When  $k = 2$  and  $k' = 1$ , the gradient  $\nabla r_h$  is constant in each element  $K_i$ , therefore,

$$\int_{\Omega} \nabla r_h \cdot w_h \, dx = \sum_{i=1}^n \nabla r_h(K_i) \cdot \int_{K_i} w_h \, dx.$$

Now the quadrature formula of exercise 6.3.7 tells us that, for  $w_h \in \mathbb{P}_2$ ,

$$\int_K w_h dx = \frac{|K|}{N(N+1)/2} \sum_j w_h(a_{ij})$$

where the  $(a_{ij})$  are the  $N(N+1)/2$  midpoints of the edges linking the vertices  $a_i$  and  $a_j$  of  $K$ . Rearranging the sum over these midpoints (which are common to two elements), we obtain

$$\int_\Omega \nabla r_h \cdot w_h dx = \sum_{a_{ij}} w_h(a_{ij}) \cdot \left( \frac{|K_i|}{N(N+1)/2} \nabla r_h(K_i) + \frac{|K_j|}{N(N+1)/2} \nabla r_h(K_j) \right).$$

Taking  $w_h$  which is 1 at the midpoint  $a_{ij}$  and 0 elsewhere, we deduce that  $W_h \cdot B_h^* R_h = 0$  implies that

$$|K_i| \nabla r_h(K_i) + |K_j| \nabla r_h(K_j) = 0. \quad (6.71)$$

The function  $r_h$  therefore has a gradient with constant direction but whose orientation changes sense from one element to the other. Since  $r_h$  belongs to  $\mathbb{P}_1$  and is continuous over  $\Omega$ , its tangential gradient is continuous at the interface between two elements. Consequently, it is zero, and the only possibility in (6.71) is that the gradient of  $r_h$  is zero everywhere, that is,  $r_h$  is a constant function. In conclusion, we have shown that  $W_h \cdot B_h^* R_h = 0$  for all  $W_h$  implies that  $R_h$  is proportional to  $\mathbb{1}$ , which is the result sought.

Let us give a counterexample in two space dimensions when  $k = k' = 1$ . We take again the uniform triangular mesh of the square  $\Omega = ]0, 1[^2$  (see Figure 6.12). We define the function  $p_0 \in Q_h$ , with  $k' = 1$ , by its values  $-1, 0, +1$  at the three vertices of each triangle  $K_i$  (see Figure 6.23). By definition, we have

$$B_h W_h \cdot R_h = \int_\Omega r_h \operatorname{div} w_h dx.$$

But as  $w_h$  is piecewise affine on each element  $K_i$ , its divergence is constant in each  $K_i$  and we have

$$\int_\Omega r_h \operatorname{div} w_h dx = \sum_{i=1}^n \operatorname{div} w_h(K_i) \int_{K_i} r_h dx$$

which is zero for  $r_h = p_0$  because  $\int_{K_i} p_0 dx = \frac{|K_i|}{3}(0 + 1 - 1) = 0$ . Consequently,  $p_0$  generates a new vector of  $\operatorname{Ker} B_h^*$ , as well as  $\mathbb{1}$ .  $\square$

In practice, **if the dimension of  $\operatorname{Ker} B_h^*$  is strictly larger than one, the corresponding finite element method is unusable.** In effect, if  $\dim(\operatorname{Ker} B_h^*) > 1$ , the numerical calculation of the solutions of the linear system (6.70) lead to numerical oscillations of the pressure: the algorithm cannot choose between several discrete pressures  $P_h$  whose difference belongs to  $\operatorname{Ker} B_h^*$ . We say that the method is **unstable**. We remark precisely that the element  $p_0$  in the proof of lemma 6.3.34 is interpreted as an oscillation of the pressure on the scale of the mesh. If  $\dim(\operatorname{Ker} B_h^*) = 1$ , we



easily eliminate the indeterminacy of the discrete pressure  $P_h$  by imposing a value at a node, or by specifying its average on the domain  $\Omega$ . In each case, there is no indeterminacy of the velocity  $U_h$  which is defined uniquely. For more details on finite element methods in fluid mechanics, we refer to [32].

We have not discussed, for the moment, the practical solution of the linear system (6.70). For this we use Uzawa's algorithm, from the theory of the optimization, which we shall see in Chapter 10. It is a very beautiful example of the interaction between numerical analysis and optimization. Let us briefly explain the principal idea (to which we return in detail in Chapter 10). We know that the Stokes equations are equivalent to a problem of minimization of an energy (see exercise 5.3.10). We see that, in the same way, the solution of the linear system (6.70) is equivalent to the following minimization

$$J(U_h) = \min_{V_h \in \text{Ker } B_h} J(V_h) \quad \text{with } J(V_h) = \frac{1}{2} A_h V_h \cdot V_h - b_h \cdot V_h.$$

Uzawa's algorithm allows us to solve exactly this minimization problem with constraints.

Due to the cost of calculation, we very rarely use the  $\mathbb{P}_2$  finite element method for the velocity and  $\mathbb{P}_1$  for the pressure. We prefer another method, called  $\mathbb{P}_1/\text{bubble}$  for the velocity and  $\mathbb{P}_1$  for the pressure. This is a  $\mathbb{P}_1$  finite element method for the velocity and the pressure in which we enrich the space  $V_{0h}$  of the velocities by adding to its basis, for each element and for each component in  $\mathbb{R}^N$ , a **bubble** function defined as the product  $\lambda_1(x) \cdots \lambda_{N+1}(x)$ , where the  $\lambda_j(x)$  are the barycentric coordinates of  $x$  in the element  $K_i$ . As this bubble function is zero on the boundary of  $K_i$  and positive in the interior, we associate it with a degree of freedom at the barycentre of the element. This method is stable as the following exercise shows.

**Exercise 6.3.20** Show that for the finite element method which is  $\mathbb{P}_1/\text{bubble}$  for the velocity and  $\mathbb{P}_1$  for the pressure we have  $\dim(\text{Ker } B_h^*) = 1$ .

The pressure instabilities are not restricted to finite element methods. There are also finite difference methods which have the same kind of disadvantage, as the following exercise shows.

**Exercise 6.3.21** We consider the Stokes equations (6.65) in  $N = 1$  dimension (this model has no interest since its explicit solution is  $u = 0$  and  $p$  is a primitive of  $f$ , but it allows us to understand the discretization problems). For  $\Omega = (0, 1)$ , we consider the mesh of points  $x_j = jh$  with  $h = 1/(n+1)$  and  $0 \leq j \leq n+1$ . We define the centred finite difference method (of order 2) as the following

$$\begin{cases} \mu \frac{-u_{j+1} + 2u_j - u_{j-1}}{h^2} + \frac{p_{j+1} - p_{j-1}}{2h} = f(x_j) & \text{for } 1 \leq j \leq n \\ \frac{u_{j+1} - u_{j-1}}{2h} = 0 & \text{for } 1 \leq j \leq n \\ u_0 = u_{n+1} = 0. \end{cases}$$

Show that this system of algebraic equations is ill-posed, and in particular that the pressure  $(p_j)$  is defined up to the addition of a constant or of a multiple of a pressure defined by the components  $(1, 0, 1, 0, \dots, 1, 0)$ .

**Remark 6.3.35** The idea of the variational formulation (6.68) extends without problem to the Laplacian or to any elliptic operator. To solve

$$\begin{cases} -\operatorname{div}(A\nabla u) = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

by setting  $\sigma = A\nabla u$ , we introduce the variational formulation

$$\begin{cases} -\int_{\Omega} \operatorname{div} \sigma v \, dx = \int_{\Omega} f v \, dx \\ \int_{\Omega} A^{-1} \sigma \cdot \tau \, dx + \int_{\Omega} u \operatorname{div} \tau \, dx = 0, \end{cases}$$

for all  $(v, \tau) \in L^2(\Omega) \times H(\operatorname{div})$ . The finite element method which follows is different from those that have been described in this chapter. It is called the mixed finite element method. •

### 6.3.5 Visualization of the numerical results

In this section we quickly say several words on the **visualization** of the results obtained by the finite element method. The figures below have been drawn with the help of the (free) graphical software xd3d.

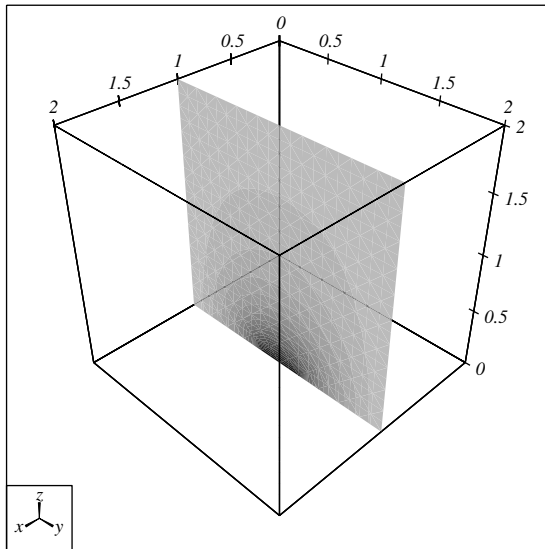


Figure 6.24. Isovalues in a cross section of a three-dimensional diffusion problem.

The visualization of the results for a scalar problem (where the unknown has values in  $\mathbb{R}$ ) is simple enough. In  $N = 2$  dimensions we can draw the isovalues and/or shade the intensity, as we can see in Figures 6.15 and 6.16. We must, nevertheless, pay attention to the scales of values as Figure 6.15 proves where the two drawings (corresponding to the same problem on two distinct meshes) are apparently comparable but the scales of reference are very different. In  $N = 3$  dimensions, we draw the isosurfaces (surfaces where the unknown is constant) or use cross sections.

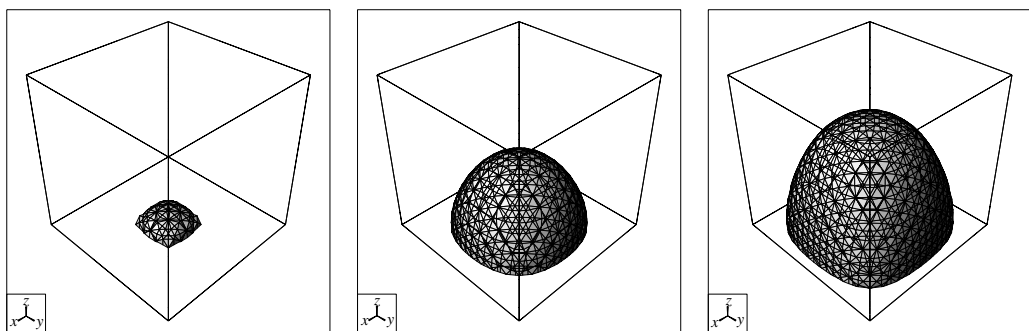


Figure 6.25. Isosurfaces for a three-dimensional diffusion (values decreasing from left to right).

As an example, we consider a diffusion problem in a concentration space (a pollutant, for example) emitted by a source localized on the ground (the base of the cube). We solve, by the  $Q_1$  finite element method, the Dirichlet problem (6.36) with a zero right-hand side and Dirichlet boundary conditions everywhere except in the base of the cube. At the centre of the base we impose the ‘nonhomogeneous’ Dirichlet boundary condition  $u = 1$ , and on the rest of the base a Neumann boundary condition. Figure 6.24 represents the values of  $u$  in a cross section, and Figure 6.25 those of the isosurfaces of  $u$ .

The visualization of the results for a vectorial problem (where the unknown has values in  $\mathbb{R}^N$ ) is different. Let us take, for example, the case of the elasticity system (see Section 5.3.1). We can draw the arrows representing the vector calculated, but this type of image is difficult to read and to interpret. It is better to draw the ‘deformation’ of the domain by using the physical interpretation of the solution (in  $N =$  two or three dimensions). Recall that the unknown vector  $u(x)$  is the displacement of the point  $x$  under the action of the exerted forces: consequently, the corresponding point in the deformed domain is  $x + u(x)$ . We illustrate these two ways to represent the results on the example of a beam fixed at its left vertical boundary (Dirichlet boundary conditions) and free on the other boundaries (Neumann boundary conditions) subject to its own weight (the force  $f$  is a constant vertical vector); see Figure 6.26. The advantage of drawing the deformed configuration is that we can superimpose the drawing of another scalar such as the norm of the tensor of the constraints.

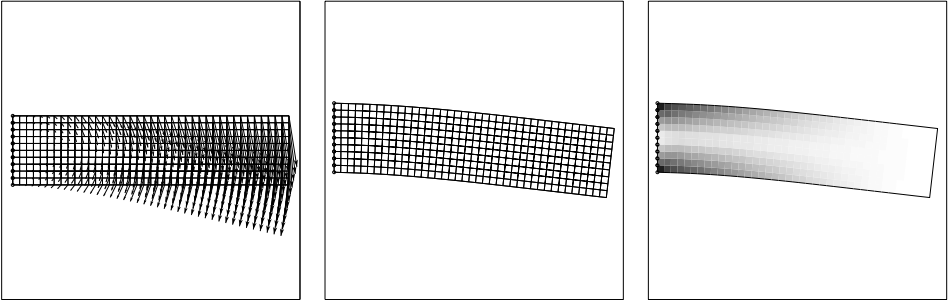


Figure 6.26. From left to right, displacement, deformed configuration, and norm of the stress tensor (the higher values are darker) in a fixed beam subject to its own weight.

*This page intentionally left blank*

# 7 Eigenvalue problems

---

## 7.1 Motivation and examples

### 7.1.1 Introduction

This chapter is dedicated to the spectral theory of partial differential equations, that is, to the study of eigenvalues and eigenfunctions of these equations. The motivation of this study is twofold. On the one hand, this will allow us to study particular solutions, which are oscillations in time (or vibrations), of the evolution problems associated with these equations. On the other hand, we shall deduce a general method for solving these evolution problems which we shall implement in Chapter 8.

Let us first give an example of an **eigenvalue problem** for the Laplacian with Dirichlet boundary conditions. If  $\Omega$  is a bounded open set of  $\mathbb{R}^N$  we look for couples  $(\lambda, u) \in \mathbb{R} \times H_0^1(\Omega)$ , with  $u \neq 0$ , which are solutions of

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega. \end{cases} \quad (7.1)$$

The real number  $\lambda$  is called the **eigenvalue**, and the function  $u(x)$  the **eigenmode or eigenfunction**. The set of eigenvalues is called the spectrum of (7.1). We can make the analogy between (7.1) and the simpler problem of determining the eigenvalues and eigenvectors of a matrix  $A$  of order  $n$ ,

$$Au = \lambda u \quad \text{with } (\lambda, u) \in \mathbb{R} \times \mathbb{R}^n, \quad (7.2)$$

confirming that the operator  $-\Delta$  is an infinite dimensional ‘generalization’ of a finite dimensional matrix  $A$ . The solution of (7.1) will be useful when solving evolution problems, either parabolic or hyperbolic, associated with the Laplacian, that is, the heat flow equation (7.5) or the wave equation (7.7). Nevertheless, the solutions of (7.1) also have a clear physical interpretation, for example, as eigenmodes of vibration.

The plan of this chapter is the following. After having motivated the eigenvalue problem (7.1) more thoroughly, we shall develop in Section 7.2 an **abstract spectral**

**theory** in Hilbert spaces. The aim of this section is to generalize to infinite dimensions the well-known result in finite dimensions which says that every real symmetric matrix is diagonalizable in an orthonormal basis. This section relies in part on a course in ‘pure’ mathematics, and we insist on the fact that it is the spirit of the results rather than the detail of the proofs that is important here. We apply this spectral theory to elliptic PDEs in Section 7.3. In particular, we show that the spectral problem (7.1) **has a countably infinite number of solutions**. Finally, Section 7.4 is dedicated to questions of **numerical approximation** of the eigenvalues and eigenfunctions of a PDE. In particular, we introduce the notion of the **mass matrix**  $\mathcal{M}$  which supplements that of the stiffness matrix  $\mathcal{K}$ , and we show that the approximate eigenvalues of (7.1) are calculated as the eigenvalues of the system  $\mathcal{K}u = \lambda\mathcal{M}u$ , which confirms the analogy between (7.1) and its discrete version (7.2).

### 7.1.2 Solution of nonstationary problems

Before launching into the abstract developments of the next section, let us show how the solution of an eigenvalue problem allows us also to solve an evolution problem. For this, we shall make an analogy with the solution of differential systems in finite dimensions. In what follows  $A$  denotes a real symmetric positive definite matrix of order  $n$ . We denote by  $\lambda_k$  its eigenvalues and  $r_k$  its eigenvectors,  $1 \leq k \leq n$ , such that  $Ar_k = \lambda_k r_k$ .

We start with a first order differential system

$$\begin{cases} \frac{\partial u}{\partial t} + Au = 0 & \text{for } t \geq 0 \\ u(t=0) = u_0, \end{cases} \quad (7.3)$$

where  $u(t)$  is a function of class  $C^1$  from  $\mathbb{R}^+$  into  $\mathbb{R}^n$ , and  $u_0 \in \mathbb{R}^n$ . It is well known that (7.3) has a unique solution obtained by diagonalizing the matrix  $A$ . More precisely, the initial data decomposes in the form  $u_0 = \sum_{k=1}^n u_k^0 r_k$ , which gives

$$u(t) = \sum_{k=1}^n u_k^0 e^{-\lambda_k t} r_k.$$

A second example is the second order differential system

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} + Au = 0 & \text{for } t \geq 0 \\ u(t=0) = u_0, \\ \frac{\partial u}{\partial t}(t=0) = u_1, \end{cases} \quad (7.4)$$

where  $u(t)$  is a function of class  $C^2$  from  $\mathbb{R}^+$  into  $\mathbb{R}^n$ , and  $u_0, u_1 \in \mathbb{R}^n$ . By decomposing the initial data in the form  $u_0 = \sum_{k=1}^n u_k^0 r_k$  and  $u_1 = \sum_{k=1}^n u_k^1 r_k$ , (7.4) has a

unique solution

$$u(t) = \sum_{k=1}^n \left( u_k^0 \cos(\sqrt{\lambda_k} t) + \frac{u_k^1}{\sqrt{\lambda_k}} \sin(\sqrt{\lambda_k} t) \right) r_k.$$

It is clear from these two examples that knowing the spectrum of the matrix  $A$  allows us to solve the evolution problems (7.3) and (7.4). These examples are representative of the path that we shall follow in the remainder of this chapter. We shall replace the matrix  $A$  by the operator  $-\Delta$ , the space  $\mathbb{R}^n$  by the Hilbert space  $L^2(\Omega)$ , and we shall ‘diagonalize’ the Laplacian to solve the heat flow equation or the wave equation.

In order to be convinced that (7.1) is the ‘good’ formulation of the eigenvalue problem for the Laplacian, we can use an argument of ‘separation of variables’ in the heat flow equation or the wave equation which we describe formally. In the absence of a source term, and by ‘forgetting’ (temporarily) the initial condition and the boundary conditions, we look for a solution  $\mathbf{u}$  of these equations which is written in the form

$$\mathbf{u}(x, t) = \phi(t)u(x),$$

that is, we separate the time and space variables. If  $\mathbf{u}$  is the solution of the heat flow equation

$$\frac{\partial \mathbf{u}}{\partial t} - \Delta \mathbf{u} = 0, \quad (7.5)$$

we find (at least formally) that

$$\frac{\phi'(t)}{\phi(t)} = \frac{\Delta u(x)}{u(x)} = -\lambda$$

where  $\lambda \in \mathbb{R}$  is a constant independent of  $t$  and of  $x$ . We deduce that  $\phi(t) = e^{-\lambda t}$  and that  $u$  must be the solution of the eigenvalue problem

$$-\Delta u = \lambda u \quad (7.6)$$

with suitable boundary conditions.

Likewise, if  $\mathbf{u}$  is the solution of the wave equation

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} - \Delta \mathbf{u} = 0, \quad (7.7)$$

we find that

$$\frac{\phi''(t)}{\phi(t)} = \frac{\Delta u(x)}{u(x)} = -\lambda$$

where  $\lambda \in \mathbb{R}$  is a constant. This time we deduce that, if  $\lambda > 0$  (which will indeed be the case), then  $\phi(t) = a \cos(\sqrt{\lambda} t) + b \sin(\sqrt{\lambda} t)$  and  $u$  must again be a solution of (7.6). Let us remark that, if the behaviour in space of the solution  $\mathbf{u}$  is the same for the heat flow equation and for the wave equation, it is not the same for its behaviour in time: it oscillates in time for the waves and decreases exponentially in time (as  $\lambda > 0$ ) for the heat.



**Remark 7.1.1** We shall see that the wave equation has solutions which oscillate periodically in time of the type

$$\mathbf{u}(x, t) = e^{-i\omega t}u(x),$$

where  $\omega = \sqrt{\lambda}$  is the frequency of the oscillations and  $u(x)$  is their amplitude. This is a general characteristic of linear hyperbolic PDEs. These oscillating solutions have an obvious physical meaning which is independent of the general solution of a hyperbolic evolution equation. They typically model vibrations (for example, elastic) or waves (for example, electromagnetic), and they occur generally in the absence of a source term and after a time which allows us to ‘forget’ the initial condition. •

**Exercise 7.1.1** Take  $\Omega = \mathbb{R}^N$ . Show that  $u(x) = \exp(ik \cdot x)$  is a solution of (7.6) if  $|k|^2 = \lambda$ . Such a solution is called a plane wave.

Let us take another example, that is, **the Schrödinger equation** from quantum mechanics (see Chapter 1).

**Exercise 7.1.2** Let  $V(x)$  be a regular potential. Show that, if  $\mathbf{u}(x, t) = e^{-i\omega t}u(x)$  is a solution of

$$i \frac{\partial \mathbf{u}}{\partial t} + \Delta \mathbf{u} - V \mathbf{u} = 0 \quad \text{in } \mathbb{R}^N \times \mathbb{R}_*^+, \quad (7.8)$$

then  $u(x)$  is a solution of

$$-\Delta u + Vu = \omega u \quad \text{in } \mathbb{R}^N. \quad (7.9)$$

We recover the same type of spectral problem as (7.6), with the addition of a zero order term. For the Schrödinger equation the eigenvalue  $\omega$  is interpreted as an energy. The smallest possible value of this energy corresponds to the energy of the fundamental state of the system described by (7.8). The other, larger, values give the energies of the excited states. Under ‘reasonable’ conditions on the potential  $V$ , these energy levels are discrete and countably infinite (which is consistent with the physical idea of *quanta*).

**Exercise 7.1.3** Take  $V(x) = Ax \cdot x$  with  $A$  a real symmetric positive definite matrix. Show that  $u(x) = \exp(-A^{1/2}x \cdot x/2)$  is a solution of (7.9) if  $\omega = \text{tr}(A^{1/2})$ . Such a solution is called a fundamental state.

## 7.2 Spectral theory

In this section we introduce an abstract spectral theory in Hilbert spaces (see for example, [25]). The ultimate aim of the developments which follow is to generalize to infinite dimensions the well-known result in finite dimensions which says that every real symmetric matrix is diagonalizable in an orthonormal basis. In first reading we can assume all the results of this section.

### 7.2.1 Generalities

In all that follows  $V$  denotes a real Hilbert space equipped with a scalar product  $\langle x, y \rangle$ .

**Definition 7.2.1** Let  $A$  be a continuous linear mapping from  $V$  into  $V$ . An eigenvalue of  $A$  is a real number  $\lambda \in \mathbb{R}$  such that there exists a nonzero element  $x \in V$  which satisfies  $Ax = \lambda x$ . Such a vector  $x$  is called the eigenvector associated with the eigenvalue  $\lambda$ .

**Theorem 7.2.2** Let  $A$  be a continuous linear mapping from  $V$  into  $V$ . There exists a unique continuous linear mapping  $A^*$  from  $V$  into  $V$ , called the adjoint, such that

$$\langle Ax, y \rangle = \langle x, A^*y \rangle \quad \forall x, y \in V.$$

**Proof.** For  $y \in V$  fixed, let  $L \in V'$  be the continuous linear form defined by  $L(x) = \langle Ax, y \rangle$ . By application of the Riesz theorem 12.1.18, there exists a unique  $z \in V$  such that  $L(x) = \langle z, x \rangle$ . We then define the mapping  $A^*$  from  $V$  into  $V$  which, to each  $y$  associates the corresponding  $z$ . We easily verify that  $A^*$  is linear and continuous, and we have  $L(x) = \langle Ax, y \rangle = \langle x, A^*y \rangle$ .  $\square$

**Definition 7.2.3** Let  $A$  be a continuous linear mapping from  $V$  into  $V$ . We say that  $A$  is self-adjoint if it coincides with its adjoint, that is  $A^* = A$ .

**Definition 7.2.4** Let  $A$  be a continuous linear mapping from  $V$  into  $V$ . We say that  $A$  is positive definite if  $\langle Ax, x \rangle > 0$  for every nonzero  $x \in V$ .

We know that in finite dimensions all self-adjoint linear mappings are diagonalizable in an orthonormal basis. We shall see that in infinite dimensions this result generalizes to continuous self-adjoint linear mappings which are in addition **compact**. Let us now introduce the ideas which allow us to define the compactness of a continuous linear mapping.

**Definition 7.2.5** A subset  $K \subset V$  is called compact if, for every sequence  $(u_n)_{n \geq 1}$  of elements of  $K$ , we can extract a subsequence  $u_{n'}$  which converges in  $K$ .

A subset  $K \subset V$  is called relatively compact if, for every sequence  $(u_n)_{n \geq 1}$  of elements of  $K$ , we can extract a subsequence  $u_{n'}$  which converges in  $V$ .

It is well known that, if  $V$  is finite dimensional, then the compact subsets of  $V$  are the closed bounded sets. Unfortunately, this result is no longer true in infinite dimensions. In effect, a compact subset is always closed and bounded but the reciprocal is not true as the following lemma shows.

**Lemma 7.2.6** In an infinite dimensional Hilbert space  $V$ , the closed unit ball is never compact.

**Proof.** As the space is infinite dimensional, we can construct, by the Gram–Schmidt procedure, an infinite orthonormal sequence  $(e_n)_{n \geq 1}$ . This sequence belongs to the closed unit ball. In addition, for  $n \neq p$  we have

$$\|e_n - e_p\|^2 = \|e_n\|^2 + \|e_p\|^2 - 2\langle e_n, e_p \rangle = 2,$$

which proves that no subsequence of  $e_n$  is a Cauchy sequence.  $\square$

**Definition 7.2.7** Let  $V$  and  $W$  be two Hilbert spaces and  $A$  a continuous linear mapping from  $V$  into  $W$ . We say that  $A$  is compact if the image under  $A$  of the unit ball of  $V$  is relatively compact in  $W$ .

Equivalently, a continuous linear mapping  $A$  is compact if, for every bounded sequence  $x_n$  of  $V$ , we can extract a subsequence such that  $Ax_{n'}$  converges in  $W$ . If  $W$  or  $V$  is finite dimensional, then every continuous linear mapping is compact. This is no longer true if  $W$  and  $V$  are infinite dimensional, as the following exercise shows.

**Exercise 7.2.1** Show that the identity mapping  $I$  in an infinite dimensional Hilbert space  $V$  is never compact (use lemma 7.2.6).

**Exercise 7.2.2** Let  $\ell_2$  be the Hilbert space of real sequences  $x = (x_i)_{i \geq 1}$  such that  $\sum_{i \geq 1} |x_i|^2 < +\infty$ , equipped with the scalar product  $\langle x, y \rangle = \sum_{i \geq 1} x_i y_i$ . Let  $(a_i)_{i \geq 1}$  be a bounded sequence of real numbers,  $|a_i| \leq C < +\infty$  for all  $i \geq 1$ . We define the linear mapping  $A$  by  $Ax = (a_i x_i)_{i \geq 1}$ . Verify that  $A$  is continuous. Show that  $A$  is compact if and only if  $\lim_{i \rightarrow +\infty} a_i = 0$ .

**Exercise 7.2.3** Let  $U$ ,  $V$ , and  $W$  be three infinite dimensional Hilbert spaces,  $A$  a continuous linear mapping from  $V$  into  $W$ , and  $B$  a continuous linear mapping from  $U$  into  $V$ . Show that the mapping  $AB$  is compact if  $A$  or  $B$  is compact. Deduce that a continuous compact linear mapping is never invertible with continuous inverse in infinite dimensions.

## 7.2.2 Spectral decomposition of a compact operator

The principal result of this section is the following.

**Theorem 7.2.8** Let  $V$  be a real infinite dimensional Hilbert space and  $A$  a continuous linear mapping which is positive definite, self-adjoint, and compact from  $V$  into  $V$ . Then the eigenvalues of  $A$  form a sequence  $(\lambda_k)_{k \geq 1}$  of strictly positive real numbers which tend to 0, and there exists a Hilbertian basis  $(u_k)_{k \geq 1}$  of  $V$  formed of eigenvectors of  $A$ , with

$$Au_k = \lambda_k u_k \quad \text{for } k \geq 1.$$

**Remark 7.2.9** As a consequence of theorem 7.2.8, and with the same notation, we obtain the **spectral decomposition** of every element  $v \in V$

$$v = \sum_{k=1}^{+\infty} \langle v, u_k \rangle u_k \quad \text{with} \quad \|v\|^2 = \sum_{k=1}^{+\infty} |\langle v, u_k \rangle|^2.$$

•

**Exercise 7.2.4** We reuse the notation and the hypotheses of theorem 7.2.8. Show that, for  $v \in V$ , the equation  $Au = v$  has a unique solution  $u \in V$  if and only if  $v$  satisfies

$$\sum_{k=1}^{+\infty} \frac{|\langle v, u_k \rangle|^2}{\lambda_k^2} < +\infty.$$

When the linear mapping  $A$  is not compact theorem 7.2.8 is false as the following exercise shows.

**Exercise 7.2.5** Take  $V = L^2(0, 1)$  and  $A$  the linear mapping from  $V$  into  $V$  defined by  $(Af)(x) = (x^2 + 1)f(x)$ . Verify that  $A$  is continuous, positive definite, self-adjoint but not compact. Show that  $A$  does not have eigenvalues. Check also that  $(A - \lambda I)$  is invertible with continuous inverse if and only if  $\lambda \notin [1, 2]$ .

To prove theorem 7.2.8 we need two preliminary lemmas.

**Lemma 7.2.10** Let  $V$  be a real Hilbert space (not containing only the zero vector) and  $A$  a continuous self-adjoint compact linear mapping from  $V$  into  $V$ . We define

$$m = \inf_{u \in V \setminus \{0\}} \frac{\langle Au, u \rangle}{\langle u, u \rangle} \quad \text{and} \quad M = \sup_{u \in V \setminus \{0\}} \frac{\langle Au, u \rangle}{\langle u, u \rangle}.$$

Then,  $\|A\| = \max(|m|, |M|)$ , and either  $m$  or  $M$  is an eigenvalue of  $A$ .

**Proof.** We easily see that  $|\langle Au, u \rangle| \leq \|A\| \|u\|^2$ , therefore  $\max(|m|, |M|) \leq \|A\|$ . On the other hand, as  $A$  is self-adjoint, we obtain for all  $u, v \in V$

$$\begin{aligned} 4\langle Au, v \rangle &= \langle A(u+v), (u+v) \rangle - \langle A(u-v), (u-v) \rangle \\ &\leq M\|u+v\|^2 - m\|u-v\|^2 \\ &\leq \max(|m|, |M|) (\|u+v\|^2 + \|u-v\|^2) \\ &\leq 2\max(|m|, |M|) (\|u\|^2 + \|v\|^2). \end{aligned}$$

Now,  $\|A\| = \sup_{\|u\|=\|v\|=1} \langle Au, v \rangle$  since  $\|Au\| = \sup_{\|v\|=1} \langle Au, v \rangle$ . We therefore deduce that  $\|A\| \leq \max(|m|, |M|)$ , from where we have the equality between these two terms.

In addition, as  $m \leq M$ , one of the following two cases holds:  $\|A\| = M \geq 0$ , or  $\|A\| = -m$  with  $m \leq 0$ . Let us consider the case  $\|A\| = M \geq 0$  (the other case  $\|A\| = -m$  is completely symmetric on replacing  $A$  by  $-A$ ). Let  $(u_n)_{n \geq 1}$  be a sequence of unit vectors of  $V$  such that

$$\lim_{n \rightarrow +\infty} \langle Au_n, u_n \rangle = M \quad \text{and} \quad \|u_n\| = 1.$$

Since  $A$  is compact, there exists a subsequence such that  $Au_{n'}$  converges in  $V$  to a limit  $v$ . On the other hand, we have

$$\langle Au_n, u_n \rangle \leq \|Au_n\| \leq \|A\| = M,$$

from which we deduce that  $\lim_{n \rightarrow +\infty} \|Au_n\| = M$ , that is,  $\|v\| = M$ . Finally, since

$$\|Au_n - Mu_n\|^2 = \|Au_n\|^2 + M^2 - 2M\langle Au_n, u_n \rangle,$$

we obtain  $\lim_{n \rightarrow +\infty} \|Au_n - Mu_n\| = 0$ . For the subsequence  $n'$ , this implies that  $u_{n'}$  converges to  $v/M$  (at least, if  $M \neq 0$ ; the case  $M = 0$  is trivial since it implies that  $A = 0$ ). By continuity of  $A$ , we therefore deduce that  $Au_{n'}$  also converges to  $Av/M$  (in addition to  $v$ ). The uniqueness of the limit shows that  $Av/M = v$ , that is,  $v$  is an eigenvector (nonzero as  $\|v\| = M \neq 0$ ) associated with the eigenvalue  $M$ .  $\square$

**Lemma 7.2.11** *Let  $V$  be a Hilbert space and  $A$  a continuous compact linear mapping from  $V$  into  $V$ . For every real number  $\delta > 0$ , there only exists a finite number of eigenvalues outside of the interval  $]-\delta, +\delta[$ , and the subspace of eigenvectors associated with each of these eigenvalues is finite dimensional.*

**Proof.** Let us show that we cannot have an infinite number of linearly independent elements of  $V$  which are eigenvectors of  $A$  for the eigenvalues  $\lambda$  such that  $|\lambda| \geq \delta > 0$ . We proceed by contradiction. Assume therefore that there exists an infinite sequence  $(u_k)_{k \geq 1}$  of elements of  $V$ , which are linearly independent, and a sequence of eigenvalues  $(\lambda_k)_{k \geq 1}$  such that

$$Au_k = \lambda_k u_k \quad \text{and} \quad |\lambda_k| \geq \delta \quad \text{for all } k \geq 1.$$

We denote by  $E_k$  the vector subspace generated by the family  $(u_1, u_2, \dots, u_k)$ . As  $E_{k-1}$  is strictly included in  $E_k$ , there exists a unit vector  $v_k \in E_k$  which is orthogonal to  $E_{k-1}$ . As  $|\lambda_k| \geq \delta$ , the sequence  $v_k/\lambda_k$  is bounded in  $V$ , and, since  $A$  is compact, we can extract a subsequence such that  $Av_{k'}/\lambda_{k'}$  converges in  $V$ . However, for  $j < k$  we can write

$$\frac{Av_k}{\lambda_k} - \frac{Av_j}{\lambda_j} = v_k + (A - \lambda_k I) \frac{v_k}{\lambda_k} - \frac{Av_j}{\lambda_j}. \quad (7.10)$$

Now, we easily verify that  $AE_k \subset E_k$  and that  $(A - \lambda_k I)E_k \subset E_{k-1}$ , therefore the two last terms on the right of (7.10) belong to  $E_{k-1}$ . As  $v_k$  is orthogonal to  $E_{k-1}$ , we deduce from (7.10)

$$\left\| \frac{Av_k}{\lambda_k} - \frac{Av_j}{\lambda_j} \right\| \geq \|v_k\| = 1,$$

which is a contradiction with the convergence of the subsequence  $Av_{k'}/\lambda_{k'}$ .  $\square$

**Proof of theorem 7.2.8.** Lemma 7.2.10 shows that the set of the eigenvalues of  $A$  is not empty, while lemma 7.2.11 shows that this set is either finite, or countably infinite with 0 as the only accumulation point. In addition, as  $A$  is positive definite, all the eigenvalues are strictly positive. Let us denote by  $(\lambda_k)$  the eigenvalues of  $A$  and  $V_k = \text{Ker}(A - \lambda_k I)$  the associated eigensubspace (lemma 7.2.11 also tells us that each  $V_k$  is finite dimensional). We remark that the eigensubspaces  $V_k$  are pairwise orthogonal: indeed, if  $v_k \in V_k$  and  $v_j \in V_j$  with  $k \neq j$ , then, as  $A$  is self-adjoint, we have

$$\langle Av_k, v_j \rangle = \lambda_k \langle v_k, v_j \rangle = \langle v_k, Av_j \rangle = \lambda_j \langle v_k, v_j \rangle,$$

from which we deduce that  $\langle v_k, v_j \rangle = 0$  since  $\lambda_k \neq \lambda_j$ . Let  $W$  be the adherence in  $V$  of the union of the  $V_k$

$$W = \overline{\left\{ u \in V, \exists K \geq 1 \text{ such that } u = \sum_{i=1}^K u_k, u_k \in V_k \right\}}.$$

We easily construct a Hilbertian basis of  $W$  by a union of the orthonormal bases of each  $V_k$  (each is finite dimensional and they are mutually orthogonal). Let us show that  $W = V$  (which will also prove that the sequence  $(\lambda_k)$  is infinite since  $V$  is infinite dimensional). We introduce the orthogonal complement of  $W$  defined by

$$W^\perp = \{u \in V \text{ such that } \langle u, v \rangle = 0 \forall v \in W\}.$$

As  $W$  is stable under  $A$  ( $AW \subset W$ ), we see that  $W^\perp$  is also stable under  $A$  because  $\langle Au, v \rangle = \langle u, Av \rangle = 0$  if  $u \in W^\perp$  and  $v \in W$ . We can therefore define the restriction of  $A$  to  $W^\perp$  which is also a continuous self-adjoint compact linear mapping. By application of lemma 7.2.10, if  $W^\perp \neq \{0\}$ , this restriction also has an eigenvalue and an eigenvector  $u \in W^\perp$  which are also an eigenvalue and eigenvector of  $A$ . This is a contradiction with the fact that, by definition,  $W$  already contains all the eigenvectors of  $A$  and that  $W \cap W^\perp = \{0\}$ . Consequently, we must have  $W^\perp = \{0\}$ , and since  $W$  is closed we deduce that  $W = \{0\}^\perp = V$ .  $\square$

**Remark 7.2.12** The proof of theorem 7.2.8 is still valid if  $A$  is not positive definite with the following restrictions: the eigenvalues are not necessarily positive, the nonzero eigenvalues can be of finite number, and  $\text{Ker} A$  (the eigensubspace associated with the zero eigenvalue) can be infinite dimensional.  $\bullet$

## 7.3 Eigenvalues of an elliptic problem

### 7.3.1 Variational problem

We return to the variational framework introduced in Chapter 3. The interest of this general framework is that it will be applied to many different models. In a Hilbert space  $V$  we consider a bilinear form  $a(\cdot, \cdot)$ , which is **symmetric**, continuous and coercive, that is,  $a(w, v) = a(v, w)$ , and there exists  $M > 0$  and  $\nu > 0$  such that

$$|a(w, v)| \leq M \|w\|_V \|v\|_V \quad \text{for all } w, v \in V$$

and

$$a(v, v) \geq \nu \|v\|_V^2 \quad \text{for all } v \in V.$$

To be able to apply the results of the preceding section, we introduce a new ingredient, that is, another Hilbert space  $H$ . We make the following fundamental hypothesis

$$\begin{cases} V \subset H \text{ with compact injection} \\ V \text{ is dense in } H. \end{cases} \quad (7.11)$$

By ‘compact injection’ we mean that the inclusion operator  $\mathcal{I}$ , which for each  $v \in V$  gives  $\mathcal{I}v = v \in H$  is continuous and compact (see definition 7.2.7). In other words,

the hypothesis (7.11) implies that from every bounded sequence of  $V$  we can extract a convergent subsequence in  $H$ . The spaces  $H$  and  $V$  do not have the same scalar product, and we denote them by  $\langle \cdot, \cdot \rangle_H$  and  $\langle \cdot, \cdot \rangle_V$  to avoid confusion.

We consider the following variational eigenvalue problem (or spectral problem): find  $\lambda \in \mathbb{R}$  and  $u \in V \setminus \{0\}$  such that

$$a(u, v) = \lambda \langle u, v \rangle_H \quad \forall v \in V. \quad (7.12)$$

We will say that  $\lambda$  is an eigenvalue of the variational problem (7.12) (or of the bilinear form  $a$ ) and that  $u$  is the associated eigenvector.

**Remark 7.3.1** Under hypothesis (7.11) the spaces  $H$  and  $V$  can never have the same scalar product. Otherwise they would be equal since  $V$  is dense in  $H$ . But this is impossible because then the injection from  $V$  into  $H$  will be the identity which is not compact (see exercise 7.2.1). •

Let us immediately give a typical concrete example of such a situation. For an open bounded set  $\Omega$ , we set  $V = H_0^1(\Omega)$ ,  $H = L^2(\Omega)$ , and the symmetric bilinear form is defined by

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx.$$

As  $C_c^\infty(\Omega)$  is dense both in  $H_0^1(\Omega)$  and  $L^2(\Omega)$ , and thanks to the Rellich theorem 4.3.21, the hypothesis (7.11) is satisfied, and we have seen in Chapter 5 that this bilinear form  $a$  is continuous and coercive over  $V$ . By a simple integration by parts, we easily see that (7.12) is equivalent to

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

that is,  $\lambda$  and  $u$  are an eigenvalue and an eigenfunction of the Laplacian.

The solutions of (7.12) are given by the following result.

**Theorem 7.3.2** *Let  $V$  and  $H$  be two real infinite dimensional Hilbert spaces. We assume that  $V \subset H$  with compact injection and that  $V$  is dense in  $H$ . Let  $a(\cdot, \cdot)$  be a symmetric bilinear form which is continuous and coercive over  $V$ . Then the eigenvalues of (7.12) form an increasing sequence  $(\lambda_k)_{k \geq 1}$  of real positive numbers which tend to infinity, and there exists a Hilbertian basis of  $H$   $(u_k)_{k \geq 1}$  of associated eigenvectors, that is,*

$$u_k \in V \quad \text{and} \quad a(u_k, v) = \lambda_k \langle u_k, v \rangle_H \quad \forall v \in V.$$

Further,  $(u_k / \sqrt{\lambda_k})_{k \geq 1}$  is a Hilbertian basis of  $V$  for the scalar product  $a(\cdot, \cdot)$ .

**Proof.** For  $f \in H$ , we solve the variational problem

$$\text{find } u \in V \text{ such that } a(u, v) = \langle f, v \rangle_H \text{ for every function } v \in V. \quad (7.13)$$

It is easy to verify the hypotheses of the Lax–Milgram theorem 3.3.1 for (7.13) which therefore has a unique solution  $u \in V$ . We define a linear mapping  $\mathcal{A}$  from  $H$  into  $V$  which for each  $f$  gives the solution  $u = \mathcal{A}f$ . In other words, the linear mapping  $\mathcal{A}$  is defined by

$$\mathcal{A}f \in V \text{ such that } a(\mathcal{A}f, v) = \langle f, v \rangle_H \text{ for all } v \in V. \quad (7.14)$$

By taking  $v = \mathcal{A}f$  in (7.14), we obtain

$$\nu \|\mathcal{A}f\|_V^2 \leq a(\mathcal{A}f, \mathcal{A}f) = \langle f, \mathcal{A}f \rangle_H \leq \|f\|_H \|\mathcal{A}f\|_H \leq C \|f\|_H \|\mathcal{A}f\|_V$$

because the injection operator  $\mathcal{I}$  from  $V$  into  $H$  is continuous. Consequently, the linear mapping  $\mathcal{A}$  is continuous from  $H$  into  $V$ . We now define a linear mapping  $A = \mathcal{I}\mathcal{A}$  from  $H$  into  $H$ , which is continuous. As  $\mathcal{I}$  is compact, the product  $A$  is also compact (see exercise 7.2.3). To show that  $A$  is self-adjoint, we take  $v = \mathcal{A}g$  in (7.14) and we obtain, for all  $f, g \in H$ ,

$$\langle f, Ag \rangle_H = \langle f, \mathcal{A}g \rangle_H = a(\mathcal{A}f, \mathcal{A}g) = a(\mathcal{A}g, \mathcal{A}f) = \langle g, \mathcal{A}f \rangle_H = \langle g, Af \rangle_H,$$

because of the symmetry of  $a$ , which proves that  $A$  is self-adjoint positive definite in  $H$ . We can therefore apply theorem 7.2.8 to the operator  $A$  which satisfies all the hypotheses. There exists a decreasing sequence  $(\mu_k)_{k \geq 1}$  of real positive numbers which tend to 0, and there exists a Hilbertian basis  $(u_k)_{k \geq 1}$  of  $H$  composed of eigenvectors of  $A$ , with

$$Au_k = \mu_k u_k \quad \text{for } k \geq 1.$$

Let us remark that, by this equality, the eigenvectors  $u_k$  belong not only to  $H$  but also to  $V$ . Let us now return to the eigenvalue problem (7.12) which can be written

$$a(u, v) = \lambda \langle u, v \rangle_H = \lambda a(\mathcal{A}u, v) \quad \forall v \in V,$$

because of the definition (7.14), that is,  $a(u - \lambda \mathcal{A}u, v) = 0$ , therefore,

$$u = \lambda \mathcal{A}u = \lambda Au.$$

Consequently, the eigenvalues  $(\lambda_k)_{k \geq 1}$  of the variational problem (7.12) are exactly the inverses of the eigenvalues  $(\mu_k)_{k \geq 1}$  of  $A$ , and their eigenvectors are the same. We set

$$\lambda_k = \frac{1}{\mu_k} \quad \text{and} \quad v_k = \frac{u_k}{\sqrt{\lambda_k}}.$$

By construction, the eigenvectors  $u_k$  form a Hilbertian basis of  $H$ . We verify that

$$a(v_k, v_j) = \frac{a(u_k, u_j)}{\sqrt{\lambda_k \lambda_j}} = \lambda_k \frac{\langle u_k, u_j \rangle_H}{\sqrt{\lambda_k \lambda_j}} = \delta_{kj},$$

and as the orthogonal complement of  $(v_k)_{k \geq 1}$  in  $V$  is contained in the orthogonal complement of  $(u_k)_{k \geq 1}$  in  $H$  (which is reduced to the zero vector), we deduce that the  $(v_k)_{k \geq 1}$  form a Hilbertian basis of  $V$  for the scalar product  $a(u, v)$ .  $\square$



**Remark 7.3.3** We insist on the fact that the operator  $\mathcal{A}$ , defined by (7.14), is the solution operator of the variational formulation, that is, it is to some extent **the inverse** of the bilinear form  $a$ . It is for this reason that the eigenvalues  $\lambda_k$  of the variational formulation are the inverses of the eigenvalues  $\mu_k$  of  $\mathcal{A}$ . For example, in finite dimensions the bilinear form is written  $a(u, v) = \mathcal{K}u \cdot v$  and we have  $\mathcal{A} = \mathcal{K}^{-1}$ . Likewise, for the Laplacian we have  $\mathcal{A} = (-\Delta)^{-1}$  (only the inverse of the Laplacian is compact, not the Laplacian itself ; see exercise 7.2.3). In fact, it is the increase in regularity of the solution of the Laplacian with respect to the right-hand side which is the source of the compactness of the operator  $(-\Delta)^{-1}$ . •

**Exercise 7.3.1** Prove a variant of theorem 7.3.2 where we replace the coercivity hypothesis on the bilinear form  $a(\cdot, \cdot)$  by the weaker hypothesis that there exists two positive constants  $\eta > 0$  and  $\nu > 0$  such that

$$a(v, v) + \eta \|v\|_H^2 \geq \nu \|v\|_V^2 \quad \text{for all } v \in V.$$

(In this case, the eigenvalues  $(\lambda_k)_{k \geq 1}$  are not strictly positive, but only satisfy  $\lambda_k + \eta > 0$ .)

In passing, we give a very useful characterization of the eigenvalues of the variational problem (7.12), called the **minimax principle or the Courant–Fisher condition**. For this we introduce the Rayleigh quotient defined, for each function  $v \in V \setminus \{0\}$ , by

$$R(v) = \frac{a(v, v)}{\|v\|_H^2}.$$

**Proposition 7.3.4 (Courant–Fisher)** *Let  $V$  and  $H$  be two real infinite dimensional Hilbert spaces. We assume that  $V \subset H$  with compact injection and that  $V$  is dense in  $H$ . Let  $a(\cdot, \cdot)$  be a bilinear form which is symmetric continuous and coercive over  $V$ . For  $k \geq 0$  we denote by  $\mathcal{E}_k$  the set of the vector subspaces of dimension  $k$  of  $V$ . We denote by  $(\lambda_k)_{k \geq 1}$  the **increasing** sequence of eigenvalues of the variational problem (7.12). Then, for all  $k \geq 1$ , the  $k$ th eigenvalue is given by*

$$\lambda_k = \min_{W \in \mathcal{E}_k} \left( \max_{v \in W \setminus \{0\}} R(v) \right) = \max_{W \in \mathcal{E}_{k-1}} \left( \min_{v \in W^\perp \setminus \{0\}} R(v) \right). \quad (7.15)$$

*In particular, the first eigenvalue satisfies*

$$\lambda_1 = \min_{v \in V \setminus \{0\}} R(v), \quad (7.16)$$

*and every minimum in (7.16) is an eigenvector associated with  $\lambda_1$ .*

**Proof.** Let  $(u_k)_{k \geq 1}$  be the Hilbertian basis of  $H$  formed by the eigenvectors of (7.12). From theorem 7.2.8,  $(u_k / \sqrt{\lambda_k})_{k \geq 1}$  is a Hilbertian basis of  $V$ . We can therefore

characterize the spaces  $H$  and  $V$  by their spectral decomposition (see remark 7.2.9)

$$H = \left\{ v = \sum_{k=1}^{+\infty} \alpha_k u_k, \quad \|v\|_H^2 = \sum_{k=1}^{+\infty} \alpha_k^2 < +\infty \right\},$$

$$V = \left\{ v = \sum_{k=1}^{+\infty} \alpha_k u_k, \quad \|v\|_V^2 = \sum_{k=1}^{+\infty} \lambda_k \alpha_k^2 < +\infty \right\}.$$

We remark in passing that, as the eigenvalues  $\lambda_k$  are bounded below by  $\lambda_1 > 0$ , this characterization shows that  $V$  is a subspace of  $H$ . We can then rewrite the Rayleigh quotient

$$R(v) = \frac{\sum_{k=1}^{+\infty} \lambda_k \alpha_k^2}{\sum_{k=1}^{+\infty} \alpha_k^2},$$

which immediately proves the result for the first eigenvalue. We introduce the subspace  $W_k \in \mathcal{E}_k$  generated by  $(u_1, u_2, \dots, u_k)$ . We have

$$R(v) = \frac{\sum_{j=1}^k \lambda_j \alpha_j^2}{\sum_{j=1}^k \alpha_j^2} \quad \forall v \in W_k \quad \text{and} \quad R(v) = \frac{\sum_{j=k}^{+\infty} \lambda_j \alpha_j^2}{\sum_{j=k}^{+\infty} \alpha_j^2} \quad \forall v \in W_{k-1}^\perp,$$

from which we deduce

$$\lambda_k = \max_{v \in W_k \setminus \{0\}} R(v) = \min_{v \in W_{k-1}^\perp \setminus \{0\}} R(v).$$

Let  $W$  be an arbitrary subspace in  $\mathcal{E}_k$ . As  $W$  has dimension  $k$  and  $W_{k-1}$  has dimension  $k-1$ , the intersection  $W \cap W_{k-1}^\perp$  is not reduced to  $\{0\}$ . Consequently,

$$\max_{v \in W \setminus \{0\}} R(v) \geq \max_{v \in W \cap W_{k-1}^\perp \setminus \{0\}} R(v) \geq \min_{v \in W \cap W_{k-1}^\perp \setminus \{0\}} R(v) \geq \min_{v \in W_{k-1}^\perp \setminus \{0\}} R(v) = \lambda_k,$$

which proves the first equality in (7.15). Likewise, if  $W$  is a subspace of  $\mathcal{E}_{k-1}$ , then  $W^\perp \cap W_k$  is not reduced to  $\{0\}$ , and

$$\min_{v \in W^\perp \setminus \{0\}} R(v) \leq \min_{v \in W^\perp \cap W_k \setminus \{0\}} R(v) \leq \max_{v \in W^\perp \cap W_k \setminus \{0\}} R(v) \leq \max_{v \in W_k \setminus \{0\}} R(v) = \lambda_k,$$

which proves the second equality in (7.15). Let  $u$  now be a minimum in (7.16). For  $v \in V$ , we introduce the function  $f(t) = R(u + tv)$  of a real variable  $t \in \mathbb{R}$  which has a minimum at  $t = 0$ . Consequently, its derivative is zero at  $t = 0$ . By taking account of the fact that  $f(0) = \lambda_1$ , a simple calculation shows that

$$f'(0) = 2 \frac{a(u, v) - \lambda_1 \langle u, v \rangle_H}{\|u\|_H^2}.$$

Since  $v$  is arbitrary in  $V$ , the condition  $f'(0) = 0$  is none other than the variational formulation (7.12), that is,  $u$  is an eigenvector associated with the eigenvalue  $\lambda_1$ .  $\square$

### 7.3.2 Eigenvalues of the Laplacian

We can immediately apply theorem 7.2.8 to the variational formulation of the Laplacian with Dirichlet boundary conditions, which gives us the following result.

**Theorem 7.3.5** *Let  $\Omega$  be a regular open bounded set, of class  $\mathcal{C}^1$ , of  $\mathbb{R}^N$ . There exists an increasing sequence  $(\lambda_k)_{k \geq 1}$  of real positive numbers which tend to infinity, and there exists a Hilbertian basis of  $L^2(\Omega)$   $(u_k)_{k \geq 1}$ , such that each  $u_k$  belongs to  $H_0^1(\Omega)$  and satisfies*

$$\begin{cases} -\Delta u_k = \lambda_k u_k & \text{a.e. in } \Omega \\ u_k = 0 & \text{a.e. on } \partial\Omega. \end{cases} \quad (7.17)$$

**Proof.** For the Laplacian with Dirichlet boundary conditions, we choose  $V = H_0^1(\Omega)$ ,  $H = L^2(\Omega)$ , and the symmetric bilinear form is defined by

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx,$$

and the scalar product over  $L^2(\Omega)$  is

$$\langle u, v \rangle_H = \int_{\Omega} uv \, dx.$$

We easily verify the hypotheses of theorem 7.2.8. Thanks to the Rellich theorem 4.3.21,  $V$  is compactly included in  $H$ . As  $C_c^\infty(\Omega)$  is dense both in  $H$  and in  $V$ ,  $V$  is dense in  $H$ . Finally, we have seen in Chapter 5 that the bilinear form  $a$  is continuous and coercive over  $V$ . Consequently, there exists an increasing sequence  $(\lambda_k)_{k \geq 1}$  of real positive numbers which tend to infinity, and there exists a Hilbertian basis of  $L^2(\Omega)$   $(u_k)_{k \geq 1}$ , such that  $u_k \in H_0^1(\Omega)$  and

$$\int_{\Omega} \nabla u_k \cdot \nabla v \, dx = \lambda_k \int_{\Omega} u_k v \, dx \quad \forall v \in H_0^1(\Omega).$$

By a simple integration by parts (of the same type that we used in the proof of theorem 5.2.2) we obtain (7.17). Let us remark that we only use the regularity of  $\Omega$  to be able to apply the trace theorem 4.3.13 and to give a meaning ‘almost everywhere’ to the Dirichlet boundary condition.  $\square$

**Remark 7.3.6** The hypothesis that the open set  $\Omega$  is bounded is absolutely fundamental in theorem 7.3.5. If it is not satisfied, the Rellich theorem 4.3.21 (about the compact injection of  $H^1(\Omega)$  into  $L^2(\Omega)$ ) is in general false, and we can show that theorem 7.3.5 no longer holds. In fact, it may be that there exists an (uncountably) infinite number of ‘generalized’ eigenvalues in the sense that the eigenfunctions do not belong to  $L^2(\Omega)$ . In the light of exercise 7.1.1 the reader can consider the case of the Laplacian in  $\Omega = \mathbb{R}^N$ .  $\bullet$

**Exercise 7.3.2** In  $N = 1$  dimension, we consider  $\Omega = ]0, 1[$ . Explicitly calculate all the eigenvalues and eigenfunctions of the Laplacian with Dirichlet boundary conditions (7.17). With the help of the spectral decomposition of this problem (see remark 7.2.9), show that the series

$$\sum_{k=1}^{+\infty} a_k \sin(k\pi x)$$

converges in  $L^2(0, 1)$  if and only if  $\sum_{k=1}^{+\infty} a_k^2 < +\infty$ , and in  $H^1(0, 1)$  if and only if  $\sum_{k=1}^{+\infty} k^2 a_k^2 < +\infty$ .

**Exercise 7.3.3** We consider a parallelepiped  $\Omega = ]0, L_1[ \times ]0, L_2[ \times \cdots \times ]0, L_N[$ , where  $(L_i > 0)_{1 \leq i \leq N}$  are positive constants. Explicitly calculate all the eigenvalues and the eigenfunctions of the Laplacian with Dirichlet boundary conditions (7.17).

Theorem 7.3.5 easily generalizes to the case of other boundary conditions. We leave the reader the task of proving the following corollary.

**Corollary 7.3.7** Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$  whose boundary  $\partial\Omega$  decomposes into two disjoint regular parts  $\partial\Omega_N$  and  $\partial\Omega_D$  (see Figure 4.1). There exists an increasing sequence  $(\lambda_k)_{k \geq 1}$  of real positive or zero numbers which tend to infinity, and there exists a Hilbertian basis of  $L^2(\Omega)$   $(u_k)_{k \geq 1}$ , such that each  $u_k$  belongs to  $H^1(\Omega)$  and satisfies

$$\begin{cases} -\Delta u_k = \lambda_k u_k & \text{in } \Omega \\ u_k = 0 & \text{on } \partial\Omega_D \\ \frac{\partial u_k}{\partial n} = 0 & \text{on } \partial\Omega_N. \end{cases}$$

**Remark 7.3.8** In the case of a purely Neumann boundary condition, that is,  $\partial\Omega_D = \emptyset$ , the bilinear form is no longer coercive over  $H^1(\Omega)$ . To prove corollary 7.3.7 we must then use exercise 7.3.1. •

**Exercise 7.3.4** We consider again an open parallelepiped  $\Omega$  as in exercise 7.3.3. Explicitly calculate all the eigenvalues and eigenfunctions of the Laplacian with Neumann boundary conditions on the entire boundary  $\partial\Omega$ .

The characterization of the eigenvalues by the Courant–Fisher principle is often very useful, as the following exercise shows.

**Exercise 7.3.5** Use again the notation and hypotheses of theorem 7.3.5. Show that the best (that is, the smallest) constant  $C$  in the Poincaré inequality (see proposition 4.3.10) is exactly the first eigenvalue  $\lambda_1$  of (7.17).

We can also show that the eigenfunctions of the Laplacian, with Dirichlet or Neumann boundary conditions, are regular.

**Proposition 7.3.9** *Let  $\Omega$  be a regular open bounded set of class  $C^\infty$ . Then the eigenfunctions which are solutions of (7.17) belong to  $C^\infty(\bar{\Omega})$ .*

**Proof.** Let  $u_k$  be the  $k$ th eigenfunction in  $H_0^1(\Omega)$  from (7.17). We can think of  $u_k$  being a solution of the following boundary value problem

$$\begin{cases} -\Delta u_k = f_k & \text{in } \Omega \\ u_k = 0 & \text{on } \partial\Omega, \end{cases}$$

with  $f_k = \lambda_k u_k$ . As  $f_k$  belongs to  $H^1(\Omega)$ , by application of the regularity theorem 5.2.26 we deduce that the solution  $u_k$  belongs to  $H^3(\Omega)$ . But in fact the right-hand side  $f_k$  is more regular which allows us again to increase the regularity of  $u_k$ . By an easy recurrence we thus show that  $u_k$  belongs to  $H^m(\Omega)$  for every  $m \geq 1$ . From theorem 4.3.25 on the continuity of  $H^m(\Omega)$  functions (see also remark 4.3.26), we deduce that  $u_k$  therefore belongs to  $C^\infty(\bar{\Omega})$ .  $\square$

We now prove a very important qualitative result with regard to the first eigenvalue.

**Theorem 7.3.10 (Krein–Rutman)** *We again take the notation and the hypotheses of theorem 7.3.5. We assume that the open set  $\Omega$  is connected. Then the first eigenvalue  $\lambda_1$  is simple (that is, the corresponding eigensubspace has dimension 1) and the first eigenvector can be chosen positive almost everywhere in  $\Omega$ .*

**Remark 7.3.11** The Krein–Rutman theorem 7.3.10 is specific to the case of ‘scalar’ equations (that is, the unknown  $u$  has values in  $\mathbb{R}$ ). This result is in general false if the unknown  $u$  is vector valued (see later the example of the elasticity system). The reason for this difference between the scalar and vector case is that this theorem relies on the maximum principle (see theorem 5.2.22) which is only valid in the scalar case.  $\bullet$

**Proof.** Let  $u \in H_0^1(\Omega)$  be a nonzero eigenvector associated with the first eigenvalue  $\lambda_1$ . From lemma 5.2.24 we know that  $u^+ = \max(u, 0)$  belongs to  $H_0^1(\Omega)$  and  $\nabla u^+ = 1_{u>0} \nabla u$  (likewise for  $u^- = \min(u, 0)$ ). Consequently, the function  $|u| = u^+ - u^-$  belongs to  $H_0^1(\Omega)$  and we have  $\nabla |u| = \text{sign}(u) \nabla u$ . Due to the Courant–Fisher proposition 7.3.4 we have

$$\lambda_1 = \min_{v \in H_0^1(\Omega) \setminus \{0\}} \left\{ R(v) \equiv \frac{\int_{\Omega} |\nabla v|^2 dx}{\int_{\Omega} v^2 dx} \right\},$$

and every minimum point is an eigenvector. Now  $\lambda_1 = R(u) = R(|u|)$ , therefore  $|u|$  is also an eigenvector associated with  $\lambda_1$ . As  $u^+$  and  $u^-$  are linear combinations of  $u$  and  $|u|$ , they are eigenfunctions associated with  $\lambda_1$ .

In fact, we can show that  $u$  is not annihilated in  $\Omega$  thanks to the ‘strong’ maximum principle which says that, if  $w \in C^2(\bar{\Omega})$  satisfies

$$-\Delta w \geq 0 \quad \text{in } \Omega \quad \text{and} \quad w = 0 \quad \text{on } \partial\Omega,$$

then either  $w \equiv 0$  in  $\Omega$ , or  $w > 0$  in  $\Omega$ . We apply this result to  $u^+$  and  $u^-$  (which are regular because of proposition 7.3.9) which cannot both be nonzero, therefore one of the two is zero. Let us now suppose that the eigensubspace associated with  $\lambda_1$  has dimension strictly greater than 1. We can then find two orthogonal eigenfunctions  $u_1$  and  $u_2$ , that is,

$$\int_{\Omega} u_1 u_2 \, dx = 0,$$

which is impossible since they have constant sign, and are nonzero.  $\square$

**Exercise 7.3.6** Let  $\Omega$  be a regular connected open bounded set. Show that the first eigenvalue of the Laplacian in  $\Omega$  with Neumann boundary condition is zero and that it is simple.

**Remark 7.3.12** The results of this section generalize without difficulty to general second order elliptic operators, that is, to the following eigenvalue problem

$$\begin{cases} -\operatorname{div}(A\nabla u) = \lambda u & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}$$

where  $A(x)$  is a symmetric coercive matrix (see Section 5.2.3).  $\bullet$

### 7.3.3 Other models

The extension of the results of the preceding section to elliptic partial differential equations which are more complicated than the Laplacian does not pose any new conceptual problems. We briefly describe this generalization for two significant examples: the linearized elasticity system and the Stokes equations.

The equations (5.56) of linear elasticity describe the stationary regime of the following dynamic equations (very similar to the wave equation)

$$\begin{cases} \rho \frac{\partial^2 u}{\partial t^2} - \operatorname{div}(2\mu e(u) + \lambda \operatorname{tr}(e(u)) \mathbf{I}) = f & \text{in } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{on } \partial\Omega \times \mathbb{R}_*^+, \end{cases} \quad (7.18)$$

where  $\rho > 0$  is the volume density of the material and  $e(u) = (\nabla u + (\nabla u)^t)/2$ . We recall that the Lamé coefficients of the material satisfy  $\mu > 0$  and  $2\mu + N\lambda > 0$ . In the absence of exterior forces  $f$  (and not taking account of possible initial conditions) we can also look for solutions of (7.18) oscillating in time like those we have described for the wave equation in Section 7.1.2. This leads to looking for the solutions  $(\ell, u)$  of the following eigenvalue problem

$$\begin{cases} -\operatorname{div}(2\mu e(u) + \lambda \operatorname{tr}(e(u)) \mathbf{I}) = \ell u & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (7.19)$$

where  $\ell = \omega^2$  is the square of the frequency of vibration (we have changed the notation of the eigenvalue to avoid confusion with the Lamé coefficient  $\lambda$ ). In mechanics, the eigenfunction  $u$  is also called the **eigenmode of vibration**.

Following the method applied above to the Laplacian we can prove the following result (we leave the details to the reader as an exercise).

**Proposition 7.3.13** *Let  $\Omega$  be a regular open bounded set of class  $C^1$  of  $\mathbb{R}^N$ . There exists an increasing sequence  $(\ell_k)_{k \geq 1}$  of real positive numbers which tend to infinity, and there exists a Hilbertian basis of  $L^2(\Omega)^N$   $(u_k)_{k \geq 1}$ , such that each  $u_k$  belongs to  $H_0^1(\Omega)^N$  and satisfies*

$$\begin{cases} -\operatorname{div}(2\mu e(u_k) + \lambda \operatorname{tr}(e(u_k)) \mathbf{I}) = \ell_k u_k & \text{a.e. in } \Omega \\ u_k = 0 & \text{a.e. on } \partial\Omega. \end{cases}$$

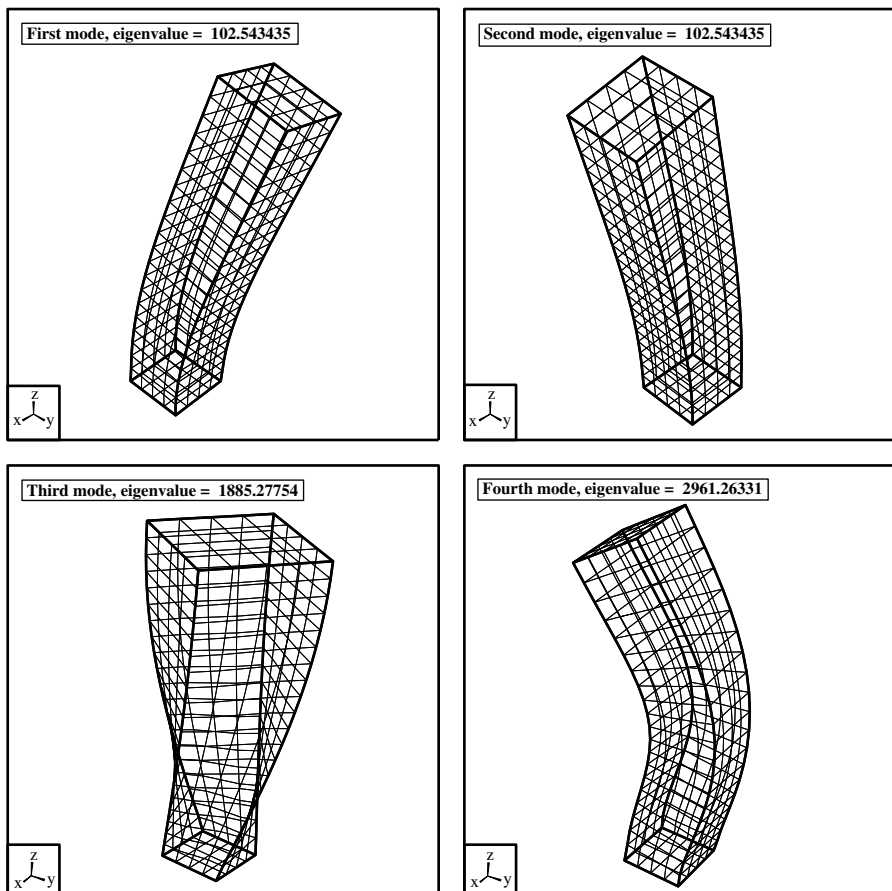


Figure 7.1. The four first eigenmodes of a 'tower' in elasticity.

The regularity result on the eigenfunctions  $u_k$  of proposition 7.3.9 also extends easily to the case of elasticity and to the problem (7.19). Conversely, theorem 7.3.10

on the simplicity of the first eigenvalue and the positivity of the first eigenfunction is in general false (as is the maximum principle). For example, we calculate by  $\mathbb{Q}_1$  finite elements the four first eigenmodes of a ‘tower’ where the base is fixed (Dirichlet boundary conditions) and the other boundaries are free (Neumann boundary condition). The two first modes correspond to the same eigenvalue (they are independent but symmetric by rotation through  $90^\circ$  around the  $z$ -axis) (see Figure 7.1 and Table 7.1). The first eigenvalue is therefore ‘double’.

Eigenmode	1	2	3	4
Eigenvalue	102.54	102.54	1885.2	2961.2

Table 7.1. Eigenvalues corresponding to the eigenmodes of Figure 7.1

We now consider the Stokes equations (5.71) which are a stationary version of a parabolic evolution problem (see later (8.2)). To solve this evolution problem it will be interesting to use the eigenvalues and eigenfunctions  $(\lambda, u, p)$  of the following problem

$$\begin{cases} \nabla p - \mu \Delta u = \lambda u & \text{in } \Omega \\ \operatorname{div} u = 0 & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (7.20)$$

where  $\mu > 0$  is the viscosity,  $u$  the velocity, and  $p$  the pressure of the fluid. By following the method applied above to the Laplacian, the reader can solve the following exercise.

**Exercise 7.3.7** Let  $\Omega$  be a regular connected open bounded set of class  $\mathcal{C}^1$  of  $\mathbb{R}^N$ . Show that there exists an increasing sequence  $(\lambda_k)_{k \geq 1}$  of real positive numbers which tend to infinity, and a Hilbertian basis  $(u_k)_{k \geq 1}$  of the subspace of  $L^2(\Omega)^N$  of the functions with zero divergence, such that each  $u_k$  belongs to  $H_0^1(\Omega)^N$ , and there exists a family of pressures  $p_k \in L^2(\Omega)$  which satisfy

$$\begin{cases} \nabla p_k - \mu \Delta u_k = \lambda_k u_k & \text{a.e. in } \Omega \\ \operatorname{div} u_k = 0 & \text{a.e. in } \Omega \\ u_k = 0 & \text{a.e. on } \partial\Omega. \end{cases}$$

The regularity result on the eigenfunctions of proposition 7.3.9 also extends easily to the case of the Stokes equations (7.20). Conversely, theorem 7.3.10 on the simplicity of the first eigenvalue and the positivity of the first eigenfunction is in general false (as is the maximum principle).

**Exercise 7.3.8** We consider the eigenvalue problem for the Schrödinger equation with a quadratic potential  $a(x) = Ax \cdot x$  where  $A$  is a symmetric positive definite matrix (a model of the harmonic oscillator)

$$-\Delta u + au = \lambda u \quad \text{in } \mathbb{R}^N. \quad (7.21)$$



We define the spaces  $H = L^2(\mathbb{R}^N)$  and

$$V = \{v \in H^1(\mathbb{R}^N) \text{ such that } |x|v(x) \in L^2(\mathbb{R}^N)\}.$$

Show that  $V$  is a Hilbert space for the scalar product

$$\langle u, v \rangle_V = \int_{\mathbb{R}^N} \nabla u(x) \cdot \nabla v(x) dx + \int_{\mathbb{R}^N} |x|^2 u(x)v(x) dx,$$

and that the injection from  $V$  into  $H$  is compact. Deduce that there exists an increasing sequence  $(\lambda_k)_{k \geq 1}$  of real positive numbers which tend to infinity and a Hilbertian basis of  $L^2(\mathbb{R}^N)$   $(u_k)_{k \geq 1}$  which are the eigenvalues and the eigenfunctions of (7.21). Explicitly calculate its eigenvalues and eigenfunctions (we shall look for  $u_k$  in the form  $p_k(x) \exp(-Ax \cdot x/2)$  where  $p_k$  is a polynomial of degree  $k - 1$ ). Interpret the results physically.

**Exercise 7.3.9** Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ . We consider the vibration problem for the plate equation with clamping boundary condition

$$\begin{cases} \Delta(\Delta u) = \lambda u & \text{in } \Omega \\ \frac{\partial u}{\partial n} = u = 0 & \text{on } \partial\Omega. \end{cases}$$

Show that there exists an increasing sequence  $(\lambda_k)_{k \geq 1}$  of positive eigenvalues which tend to infinity and a Hilbertian basis in  $L^2(\Omega)$  of eigenfunctions  $(u_k)_{k \geq 1}$  which belong to  $H_0^2(\Omega)$ .

## 7.4 Numerical methods

### 7.4.1 Discretization by finite elements

We shall consider an internal approximation of the variational formulation introduced in Section 7.3.1. Given a subspace  $V_h$  of the finite dimensional Hilbert space  $V$ , we look for the solution  $(\lambda_h, u_h) \in \mathbb{R} \times V_h$  of

$$a(u_h, v_h) = \lambda_h \langle u_h, v_h \rangle_H \quad \forall v_h \in V_h. \quad (7.22)$$

Typically,  $V_h$  is a finite element space like those introduced in definitions 6.3.5 and 6.3.25, and  $H$  is the space  $L^2(\Omega)$ . The solution of the internal approximation (7.22) is easy as the following lemma shows.

**Lemma 7.4.1** *We take the hypotheses of theorem 7.3.2. Then the eigenvalues of (7.22) form a finite increasing sequence*

$$0 < \lambda_1 \leq \dots \leq \lambda_{n_{dl}} \quad \text{with } n_{dl} = \dim V_h,$$

*and there exists a basis of  $V_h$ , which is orthonormal in  $H$ ,  $(u_{k,h})_{1 \leq k \leq n_{dl}}$  of associated eigenvectors, that is,*

$$u_{k,h} \in V_h, \text{ and } a(u_{k,h}, v_h) = \lambda_k \langle u_{k,h}, v_h \rangle_H \quad \forall v_h \in V_h.$$

**Proof.** This can be considered as an obvious variant of theorem 7.3.2 (up to the difference that in finite dimensions there exist a finite number of eigenvalues). Nevertheless, we give a different proof, which is purely algebraic, which corresponds more closely to the steps followed in practice. Let  $(\phi_i)_{1 \leq i \leq n_{dl}}$  be a basis of  $V_h$  (for example, the finite element basis functions, see proposition 6.3.7). We look for  $u_h$  a solution of (7.22) in the form

$$u_h(x) = \sum_{i=1}^{n_{dl}} U_i^h \phi_i(x).$$

Introducing the **mass matrix**  $\mathcal{M}_h$  defined by

$$(\mathcal{M}_h)_{ij} = \langle \phi_i, \phi_j \rangle_H \quad 1 \leq i, j \leq n_{dl},$$

and the **stiffness matrix**  $\mathcal{K}_h$  defined by

$$(\mathcal{K}_h)_{ij} = a(\phi_i, \phi_j) \quad 1 \leq i, j \leq n_{dl},$$

the problem (7.22) is equivalent to finding  $(\lambda_h, U_h) \in \mathbb{R} \times \mathbb{R}^{n_{dl}}$  the solution of

$$\mathcal{K}_h U_h = \lambda_h \mathcal{M}_h U_h. \quad (7.23)$$

The names ‘mass and stiffness matrices’ come from applications in solid mechanics. Let us remark that, in the case where  $V_h$  is a finite element space, the stiffness matrix  $\mathcal{K}_h$  is exactly the same matrix that we met in chapter 6 in the application of the finite element method to elliptic problems. We verify immediately that the matrices  $\mathcal{M}_h$  and  $\mathcal{K}_h$  are symmetric and positive definite. The system (7.23) is a ‘generalised’ matrix eigenvalue problem. The simultaneous reduction theorem (see, for example, [24]) confirms that there exists an invertible matrix  $P_h$  such that

$$\mathcal{M}_h = P_h P_h^*, \quad \text{and} \quad \mathcal{K}_h = P_h \text{diag}(\lambda_k) P_h^*.$$

Consequently, the solutions of (7.23) are the eigenvalues  $(\lambda_k)$  and the eigenvectors  $(U_{k,h})_{1 \leq k \leq n_{dl}}$  which are the column vectors of  $(P_h^*)^{-1}$ . These column vectors, therefore, form a basis, which is orthogonal for  $\mathcal{K}_h$  and orthonormal for  $\mathcal{M}_h$  (we shall briefly indicate in remark 7.4.3 how to calculate this basis). Finally, the vectors  $U_{k,h}$  are simply the vectors of the coordinates in the basis  $(\phi_i)_{1 \leq i \leq n_{dl}}$  of the functions  $u_{k,h}$  which form an orthonormal basis of  $V_h$  for the scalar product of  $H$ .  $\square$

**Remark 7.4.2** In lemma 7.4.1 we have used the hypotheses of theorem 7.3.2: in particular, the bilinear form  $a(u, v)$  is assumed **symmetric**. We see the importance of this hypothesis in the proof. In effect, if it is not symmetric, we would not know if the system (7.23) is diagonalizable, that is, if there exist solutions of the eigenvalue problem (7.22).  $\bullet$

The application of lemma 7.4.1 to the variational approximation by finite elements of the Dirichlet problem (7.17) is straightforward. We take  $V = H_0^1(\Omega)$ ,  $H = L^2(\Omega)$ , and the discrete space  $V_{0h}$  of definition 6.3.5 (recall that  $V_{0h}$  contains the Dirichlet boundary conditions).

**Exercise 7.4.1** We consider the eigenvalue problem in  $N = 1$  dimension

$$\begin{cases} -u_k'' = \lambda_k u_k & \text{for } 0 < x < 1 \\ u_k(0) = u_k(1) = 0 \end{cases}.$$

We propose to calculate the mass matrix for the  $\mathbb{P}_1$  finite element method. Using the notation of Section 6.2. Show that the mass matrix  $\mathcal{M}_h$  is given by

$$\mathcal{M}_h = h \begin{pmatrix} 2/3 & 1/6 & & & 0 \\ 1/6 & 2/3 & 1/6 & & \\ & \ddots & \ddots & \ddots & \\ & & 1/6 & 2/3 & 1/6 \\ 0 & & & 1/6 & 2/3 \end{pmatrix},$$

and that its eigenvalues are

$$\lambda_k(\mathcal{M}_h) = \frac{h}{3} (2 + \cos(k\pi h)) \quad \text{for } 1 \leq k \leq n.$$

Show that, if we use the quadrature formula (6.45), then we find that  $\mathcal{M}_h = hI$ . In this last case, calculate the eigenvalues of the discrete spectral problem.

**Remark 7.4.3** To calculate the eigenvalues and eigenvectors of the spectral matrix problem (7.23) we must, in general, start by calculating the Cholesky factorization of the mass matrix  $\mathcal{M}_h = \mathcal{L}_h \mathcal{L}_h^*$ , to reduce it to the classical case

$$\tilde{\mathcal{K}}_h \tilde{U}_h = \lambda_h \tilde{U}_h \quad \text{with } \tilde{\mathcal{K}}_h = \mathcal{L}_h^{-1} \mathcal{K}_h (\mathcal{L}_h^*)^{-1} \quad \text{and} \quad \tilde{U}_h = \mathcal{L}_h^* U_h,$$

for which we have algorithms to calculate the eigenvalues and eigenvectors. We refer to Section 13.2 for more details about these algorithms: let us only say that this is the most expensive step in calculation time.

We can avoid the construction of the matrix  $\tilde{\mathcal{K}}_h$  and make the Cholesky factorization of  $\mathcal{M}_h$  less expensive if we use a quadrature formula to evaluate the coefficients of the matrix  $\mathcal{M}_h$  which makes it **diagonal**. This numerical integration procedure is called **mass lumping**, or condensation, and is frequently used. For example, if we use the quadrature formula (6.45) (which only uses the values of a function at the nodes to calculate an integral), we easily see that the mass matrix  $\mathcal{M}_h$  obtained is diagonal (see exercise 7.4.1). •

We see in the following section that **only the first discrete eigenvalues**  $\lambda_{k,h}$  (the smallest) are correct approximations to the exact eigenvalues  $\lambda_k$  (likewise for the eigenvectors). We must therefore pay attention to the fact that the last eigenvalues (the largest) of the discrete problem (7.23) do not have any physical significance! Consequently, if we are interested in the thousandth eigenvalue of the Dirichlet problem (7.17), we must use a sufficiently fine mesh of the domain  $\Omega$  so that the dimension of the finite element space  $V_h$  is much greater than a thousand.

We illustrate this section by the calculation of the six first eigenmodes of vibration of a drum (modelled as a circular membrane fixed at its boundary). We therefore solve the Dirichlet problem (7.17) in a disc of radius 1. We use a  $\mathbb{P}_1$  finite element method. The results are presented in Figure 7.2 and Table 7.2. We remark that the first eigenvalue is simple while the second and the third are ‘doubles’.

Eigenmode	1	2	3	4	5	6
Eigenvalue	5.78	14.69	14.69	26.42	26.42	30.53

Table 7.2. Eigenvalues corresponding to the eigenmodes of Figure 7.2.

### 7.4.2 Convergence and error estimates

In this section, we restrict ourselves to stating a convergence result for the  $\mathbb{P}_k$  triangular finite element method for the calculation of eigenvalues and eigenvectors of the Dirichlet problem (7.17). It is clear that this result generalizes easily to other problems and to other types of finite elements.

**Theorem 7.4.4** *Let  $(\mathcal{T}_h)_{h>0}$  be a sequence of regular triangular meshes of  $\Omega$ . Let  $V_{0h}$  be the subspace of  $H_0^1(\Omega)$ , defined by the  $\mathbb{P}_k$  finite element method, with dimension  $n_{dl}$ . Let  $(\lambda_i, u_i) \in \mathbb{R} \times H_0^1(\Omega)$ , for  $i \geq 1$ , be the eigenvalues and eigenvectors (orthonormal in  $L^2(\Omega)$ ) of the Dirichlet problem (6.36), arranged in increasing order*

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_i \leq \lambda_{i+1} \cdots$$

Let

$$0 < \lambda_{1,h} \leq \lambda_{2,h} \leq \cdots \leq \lambda_{n_{dl},h},$$

be the eigenvalues of the variational approximation (7.22) in  $V_{0h}$ . For every fixed  $i \geq 1$ , we have

$$\lim_{h \rightarrow 0} |\lambda_i - \lambda_{i,h}| = 0. \quad (7.24)$$

There exists a family of eigenvectors  $(u_{i,h})_{1 \leq i \leq n_{dl}}$  of (7.22) in  $V_{0h}$  such that, if  $\lambda_i$  is a simple eigenvalue, we have

$$\lim_{h \rightarrow 0} \|u_i - u_{i,h}\|_{H^1(\Omega)} = 0. \quad (7.25)$$

Further, if the subspace generated by  $(u_1, \dots, u_i)$  is included in  $H^{k+1}(\Omega)$  and if  $k+1 > N/2$ , then we have the error estimate

$$|\lambda_i - \lambda_{i,h}| \leq C_i h^{2k}, \quad (7.26)$$

where  $C_i$  does not depend on  $h$ , and if  $\lambda_i$  is a simple eigenvalue, we have

$$\|u_i - u_{i,h}\|_{H^1(\Omega)} \leq C_i h^k. \quad (7.27)$$

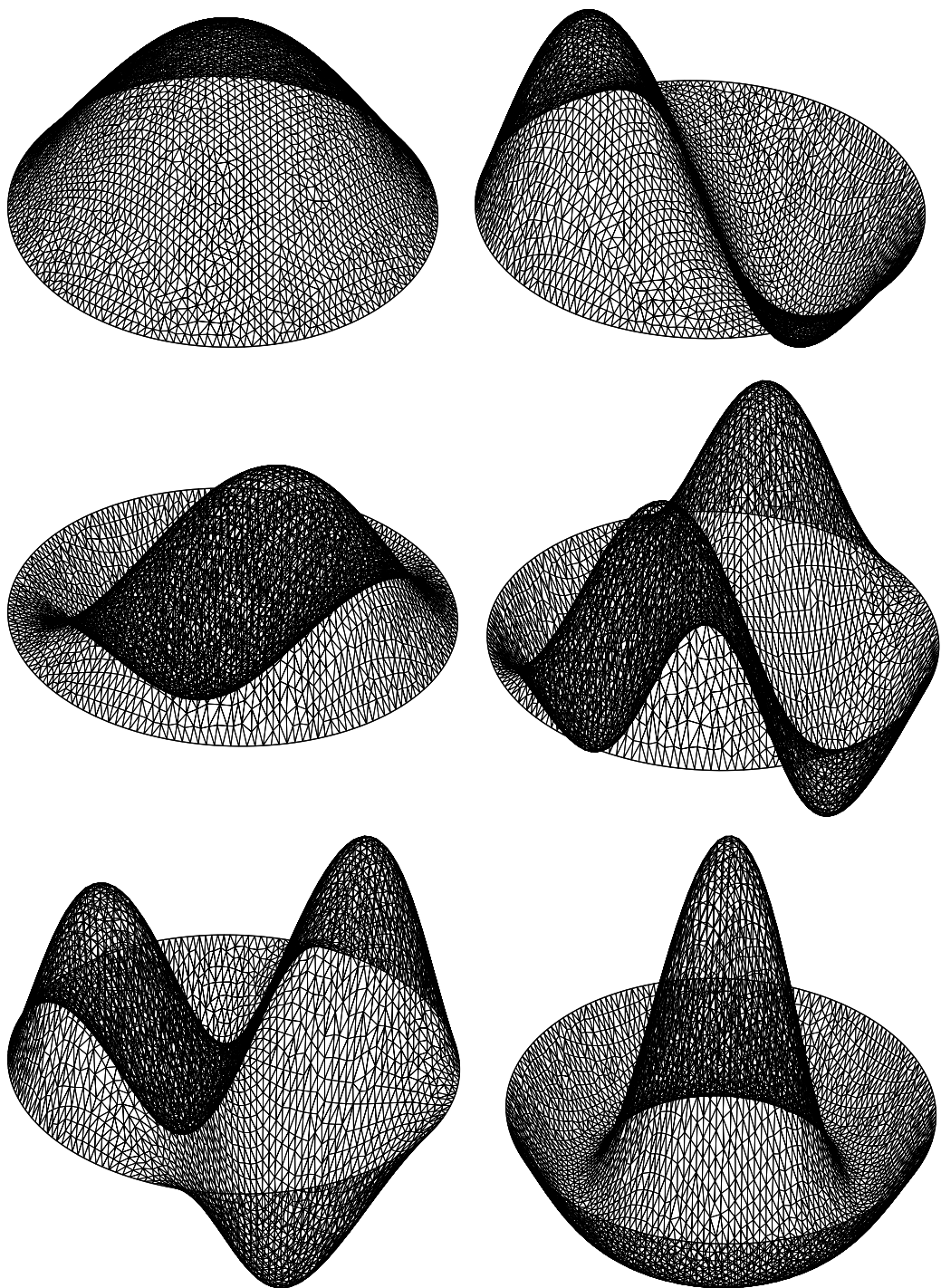


Figure 7.2. The six first eigenmodes of a drum.

**Remark 7.4.5** The constant  $C_i$  in (7.26) or (7.27) tends to  $+\infty$  as  $i \rightarrow +\infty$ , which means that there is no guarantee that the larger discrete eigenvalues (for example,  $\lambda_{n_{dl},h}$ ) approximate the exact eigenvalues  $\lambda_i$ .

The order of convergence of the eigenvalues is twice that of the convergence of the eigenvectors. This is a general phenomenon in the approximation of self-adjoint spectral operators which we shall see again in the numerical algorithms of Section 13.2 (see proposition 13.2.1).

The convergence of the eigenvectors can only be obtained if the corresponding eigenvalue is simple. In effect, if the eigenvalue  $\lambda_i$  is multiple, the sequence of approximate eigenvectors  $u_{i,h}$  cannot converge and have several accumulation points which are different linear combinations of the eigenvectors of the eigensubspace associated with  $\lambda_i$ . •

*This page intentionally left blank*

# 8 Evolution problems

## 8.1 Motivation and examples

### 8.1.1 Introduction

This chapter is dedicated to the mathematical and numerical analysis of problems of evolution in time (in the preceding chapters we have studied stationary problems without a time variable). More precisely we shall analyse two different types of partial differential equations: parabolic, and hyperbolic. The typical example of a parabolic equation is the heat flow equation which we will study in detail (but our analysis extends to more complicated models like the time-dependent Stokes equations). The prototype of a hyperbolic equation is the wave equation on which we shall concentrate (but once more our analysis extends to more complicated models like the elastodynamic equations or the electromagnetic equations). More generally, the approach developed here extends to many other evolution problems, not necessarily parabolic or hyperbolic, like, for example, the Schrödinger equation in quantum mechanics.

The plan of this chapter is the following. The remainder of this section is dedicated to some questions linked to modelling. In Sections 8.2 and 8.3 we prove the existence and uniqueness of the solution of the heat flow equation or of the wave equation by again using the concept of **variational formulation**. As we have seen in Section 7.1.2, we use the **Hilbertian bases of eigenfunctions** constructed in Chapter 7. We shall also use the idea of **energy estimates** which express a physical balance of energy and partly justifies the spaces used by the theory. In Sections 8.4 and 8.5 we shall study some **qualitative properties** of the solutions. Let us point out immediately that, while the existence and uniqueness results are very similar for the heat flow equation and the wave equation, **their qualitative properties are, conversely, very different**. We also see that, while some of these qualitative properties agree with physical intuition (like the maximum principle for heat flow, and conservation of energy for the waves), others are more surprising, and for this reason particularly interesting (like the ‘infinite’ speed of propagation of heat, and the reversibility in



time of the waves). All these results will be proved for an equation posed in a bounded domain. Nevertheless, we shall say a few words about the situation when the equation is posed in the whole space  $\mathbb{R}^N$ ; in this case, we can obtain an explicit formula for the solution by using a Green's function.

Sections 8.6 and 8.7 are dedicated to the **numerical solution** of the heat flow and wave equations. We have already explored the finite difference method for these problems (see Chapters 1 and 2). We shall concentrate on the use of the **finite element** method in this context. More precisely, as is mostly done, we use finite elements for the spatial discretization, but finite differences for the temporal discretization.

### 8.1.2 Modelling and examples of parabolic equations

Let us quickly present the principal parabolic problems that we shall study in this chapter, by saying some words about their physical or mechanical origin. The archetype of these models is the **heat flow equation** whose physical origin has already been discussed in Chapter 1. Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$  with boundary  $\partial\Omega$ . For Dirichlet boundary conditions this model is written

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{in } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{on } \partial\Omega \times \mathbb{R}_*^+ \\ u(x, 0) = u_0(x) & \text{for } x \in \Omega. \end{cases} \quad (8.1)$$

The boundary value problem (8.1) models the evolution of the temperature  $u(x, t)$  in a thermally conducting body occupying the domain  $\Omega$ . The distribution of the initial temperature, at  $t = 0$ , is given by the function  $u_0$ . On the boundary  $\partial\Omega$  of the body, the temperature is maintained at a constant value, and used as the reference value (this is the homogeneous Dirichlet condition  $u(x, t) = 0$  on  $\partial\Omega \times \mathbb{R}_+$ ). The heat sources are modelled by the given function  $f = f(x, t)$ . Let us note that the variables  $x \in \Omega$  and  $t \in \mathbb{R}_+$  play very different roles in (8.1) since it is a partial differential equation of first order in  $t$  and second order in  $x$  (the Laplacian only acts on the spatial variable).

Let us mention that there exist other physical origins of system (8.1). For example, (8.1) also models the diffusion of a concentration  $u$  in the domain  $\Omega$ , or the evolution of the pressure field  $u$  of a fluid flowing in a porous medium (Darcy flow), or even Brownian motion in the domain  $\Omega$ .

We can, of course, associate others boundary conditions with the heat flow equation (for example, a homogeneous Neumann condition if the wall of the body  $\Omega$  is adiabatic).

A first obvious generalization of the heat flow equation is obtained when we replace the Laplacian by a more general second order elliptic operator (see Section 5.2.3). This generalization is met, for example, if we study the propagation of heat in a nonhomogeneous material or in the presence of a convective effect. A second (less obvious) generalization concerns the system of time-dependent Stokes equations that we have quickly stated in the preceding chapter. Denoting by  $u$  the velocity and  $p$

the pressure of a viscous fluid subject to the force  $f$ , this system is written

$$\begin{cases} \frac{\partial u}{\partial t} + \nabla p - \mu \Delta u = f & \text{in } \Omega \times \mathbb{R}_*^+ \\ \operatorname{div} u = 0 & \text{in } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{on } \partial\Omega \times \mathbb{R}_*^+ \\ u(x, t = 0) = u_0(x) & \text{in } \Omega \end{cases} \quad (8.2)$$

where  $\mu > 0$  is the viscosity of the fluid. Let us recall that the homogeneous Dirichlet boundary condition models the adherence of the fluid to the boundary of  $\Omega$  (see Chapter 1), and that the Stokes system is only valid for slow velocities (see remark 5.3.7). Most of the results that we see in this chapter generalize to such models.

### 8.1.3 Modelling and examples of hyperbolic equations

Let us quickly present the two principal hyperbolic models that we shall study in this chapter. The first model is **the wave equation** whose physical origin has already been discussed in Chapter 1. Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$  with boundary  $\partial\Omega$ . For Dirichlet boundary conditions this model is written

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta u = f & \text{in } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{on } \partial\Omega \times \mathbb{R}_*^+ \\ u(t = 0) = u_0(x) & \text{in } \Omega \\ \frac{\partial u}{\partial t}(t = 0) = u_1(x) & \text{in } \Omega. \end{cases} \quad (8.3)$$

The boundary value problem (8.3) models, for example, the propagation in time of the vertical displacement of an elastic membrane, or the amplitude of an electric field with constant direction. The unknown  $u(t, x)$  is a scalar function here.

The second model is the **elastodynamic** model which is the time dependent version of the linearized elasticity equations (see Chapters 1 and 5). By applying the fundamental principle of dynamics, the acceleration being the second time derivative of the displacement, we obtain an evolution problem which is second order in time (8.3). Nevertheless, an important difference from (8.3) is that the unknown  $u(t, x)$  is from now on a vector valued function in  $\mathbb{R}^N$ . More precisely, if we denote by  $f(t, x)$  the (vector) resultant of the exterior forces, the displacement  $u(t, x)$  is the solution of

$$\begin{cases} \rho \frac{\partial^2 u}{\partial t^2} - \operatorname{div} (2\mu e(u) + \lambda \operatorname{tr}(e(u)) \mathbf{I}) = f & \text{in } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{on } \partial\Omega \times \mathbb{R}_*^+ \\ u(t = 0) = u_0(x) & \text{in } \Omega \\ \frac{\partial u}{\partial t}(t = 0) = u_1(x) & \text{in } \Omega, \end{cases} \quad (8.4)$$

where  $u_0$  is the initial displacement,  $u_1$  the initial velocity, and  $e(u) = (\nabla u + (\nabla u)^t)/2$  the deformation tensor. Assuming that the material which occupies  $\Omega$  is homogeneous and isotropic, its density is constant  $\rho > 0$ , as are its Lamé coefficients which satisfy  $\mu > 0$  and  $2\mu + N\lambda > 0$ .

**Remark 8.1.1** We can add a term which is first order in time to the equations (8.3) and (8.4), which gives

$$\frac{\partial^2 u}{\partial t^2} + \eta \frac{\partial u}{\partial t} - \Delta u = f \quad \text{in } \Omega \times \mathbb{R}_*^+.$$

When the coefficient  $\eta$  is positive, this first order term corresponds to a braking force proportional to the velocity. We also say that it is a damping term. We then talk of the **damped wave equation**. •

**Remark 8.1.2** There are other physical models which lead to hyperbolic partial differential equations. However, every hyperbolic model is not necessarily a second order evolution problem. This is notably the case for the linearized Euler equations in acoustics, or Maxwell's equations in electromagnetism, which are systems of hyperbolic equations which are only first order in time. The ideas contained in this chapter extend to these problems, but the different order in time changes the presentation. •

## 8.2 Existence and uniqueness in the parabolic case

We shall follow a path similar in spirit to that which guided us in Chapter 5 to establish the existence and uniqueness of the solution of an elliptic problem. This path breaks into three steps: first, to establish a variational formulation (Section 8.2.1), second, to prove the existence and uniqueness of the solution of this variational formulation by using a Hilbertian basis of eigenfunctions (Section 8.2.2), third, to show that this solution satisfies the boundary value problem studied (Section 8.2.3).

### 8.2.1 Variational formulation

The idea is to write a variational formulation which resembles a first order **ordinary differential equation**, similar to (7.3). For this we multiply the heat flow equation (8.1) by a test function  $v(x)$  which does not depend on time  $t$ . Because of the boundary condition we shall demand that  $v$  is zero on the boundary of the open set  $\Omega$ , which will allow us to carry out a simple integration by parts (without boundary term). For the moment, this calculation is formal. We therefore obtain

$$\int_{\Omega} \frac{\partial u}{\partial t}(x, t) v(x) dx + \int_{\Omega} \nabla u(x, t) \cdot \nabla v(x) dx = \int_{\Omega} f(x, t) v(x) dx. \quad (8.5)$$

Since neither  $\Omega$  nor  $v(x)$  vary with time  $t$ , we can rewrite this equation in the form

$$\frac{d}{dt} \int_{\Omega} u(x, t) v(x) dx + \int_{\Omega} \nabla u(x, t) \cdot \nabla v(x) dx = \int_{\Omega} f(x, t) v(x) dx.$$

Exploiting the fact that the variables  $x$  and  $t$  play very different roles, we separate these variables by considering from now on the solution  $u(t, x)$  as a function of time  $t$  **with values in a space of functions** defined over  $\Omega$  (likewise for  $f(t, x)$ ). More precisely, if we are given a final time  $T > 0$  (possibly equal to  $+\infty$ ), it is considered that  $u$  is defined by

$$\begin{aligned} u : ]0, T[ &\rightarrow H_0^1(\Omega) \\ t &\rightarrow u(t), \end{aligned}$$

and we continue to use the notation  $u(x, t)$  for the value  $u(t)(x)$ . The choice of the space  $H_0^1(\Omega)$  is obviously dictated by the nature of the problem and can vary from one model to another. Generally it is the space which is suitable for the variational formulation of the associated stationary problem. Likewise, the source term  $f$  is from now on considered as a function of  $t$  with values in  $L^2(\Omega)$ .

We then introduce the scalar product of  $L^2(\Omega)$  and the bilinear form  $a(w, v)$  defined by

$$\langle w, v \rangle_{L^2(\Omega)} = \int_{\Omega} w(x)v(x) dx \quad \text{and} \quad a(w, v) = \int_{\Omega} \nabla w(x) \cdot \nabla v(x) dx.$$

By choosing the test function in the space  $H_0^1(\Omega)$ , we can then put (8.5) into the form of an **ordinary differential equation** in  $t$ . We thus obtain the following variational formulation: find  $u(t)$ , a function of  $]0, T[$ , with values in  $H_0^1(\Omega)$  such that

$$\begin{cases} \frac{d}{dt} \langle u(t), v \rangle_{L^2(\Omega)} + a(u(t), v) = \langle f(t), v \rangle_{L^2(\Omega)} & \forall v \in H_0^1(\Omega), 0 < t < T, \\ u(t=0) = u_0. \end{cases} \quad (8.6)$$

There are several points to be made clear in the variational formulation (8.6) in order to give it a precise mathematical meaning: what is the regularity in time of  $f$  and  $u$ , and what meaning do we give to the derivative in time? In particular, it is absolutely necessary that  $u(t)$  is continuous at  $t = 0$  to give a correct meaning to the initial data  $u_0$ .

For this we need to introduce a family of functional spaces of functions of  $t$  with values in the spaces of functions of  $x$ .

**Definition 8.2.1** *Let  $X$  be a Hilbert space, or more generally, a Banach space defined over  $\Omega$  (typically,  $X = L^2(\Omega)$ ,  $H_0^1(\Omega)$ , or  $C(\overline{\Omega})$ ). Take a final time  $0 < T \leq +\infty$ . For an integer  $k \geq 0$ , we denote by  $C^k([0, T]; X)$  the space of functions  $k$  times continuously differentiable in  $[0, T]$  belonging to  $X$ . If we denote the norm in  $X$  by  $\|v\|_X$ , it is classical (see [28]) that  $C^k([0, T]; X)$  is a Banach space for the norm*

$$\|v\|_{C^k([0, T]; X)} = \sum_{m=0}^k \left( \sup_{0 \leq t \leq T} \left\| \frac{d^m v}{dt^m}(t) \right\|_X \right).$$

We denote by  $L^2(]0, T[; X)$  the space of functions of  $]0, T[$  in  $X$  such that the function  $t \rightarrow \|v(t)\|_X$  is measurable and square integrable, that is, to say that

$$\|v\|_{L^2(]0, T[; X)} = \sqrt{\int_0^T \|v(t)\|_X^2 dt} < +\infty.$$

Equipped with this norm  $L^2(]0, T[; X)$  is also a Banach space. Further, if  $X$  is a Hilbert space, then  $L^2(]0, T[; X)$  is a Hilbert space for the scalar product

$$\langle u, v \rangle_{L^2(]0, T[; X)} = \int_0^T \langle u(t), v(t) \rangle_X dt.$$

**Remark 8.2.2** If  $X$  is the space  $L^2(\Omega)$ , then  $L^2(]0, T[; L^2(\Omega))$  identifies with the space  $L^2(]0, T[ \times \Omega)$  since, by the Fubini theorem, we have

$$\|v\|_{L^2(]0, T[; L^2(\Omega))}^2 = \int_0^T \left( \int_{\Omega} |v(t)|^2(x) dx \right) dt = \int_0^T \int_{\Omega} |v(x, t)|^2 dx dt = \|v\|_{L^2(]0, T[ \times \Omega)}^2.$$

For  $1 \leq p < +\infty$ , we can generalize definition 8.2.1 by introducing the Banach space  $L^p(]0, T[; X)$  of functions of  $]0, T[$  in  $X$  such that the function  $t \rightarrow \|v(t)\|_X$  is measurable and has integrable  $p$ th power. •

In what follows, we shall take the source term  $f$  in the space  $L^2(]0, T[; L^2(\Omega))$ , and we shall look for the solution  $u$  in **the energy space**  $L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ . This choice may appear arbitrary, but it will be **justified**, not only by the proof of the existence of a solution for the variational formulation (8.6), but also **by its link with the energy estimates** (see exercise 8.2.1). Let us note already that, as the functions of this space are continuous in time with values in  $L^2(\Omega)$ , the initial condition has a meaning.

Finally, the time derivative in the variational formulation (8.6) must be taken in a weak sense since *a priori* the function  $t \rightarrow \langle u(t), v \rangle_{L^2(\Omega)}$  does not belong to  $L^2(0, T)$  (see lemma 4.4.12 for a precise definition of this idea of derivative). Very fortunately if there exists a solution of (8.6), then the equality in (8.6) tells us that this time derivative is all the more classical since it belongs to  $L^2(]0, T[)$ .

## 8.2.2 A general result

To prove the existence and uniqueness of the solution of the variational formulation (8.6), we return to the general framework introduced in Section 7.3. We shall therefore be able to ‘diagonalize’ the Laplacian operator and to reduce the problem to the solution of a family of simple first order ordinary differential equations. We therefore introduce two Hilbert spaces  $V$  and  $H$  such that  $V \subset H$  with dense and compact injection (see (7.11) and the explanation which follows). Typically we will have  $V = H_0^1(\Omega)$  and  $H = L^2(\Omega)$ .

**Theorem 8.2.3** *Let  $V$  and  $H$  be two Hilbert spaces such that  $V \subset H$  with compact injection and  $V$  is dense in  $H$ . Let  $a(u, v)$  be a symmetric bilinear form which is continuous and coercive in  $V$ . Take a final time  $T > 0$ , initial data  $u_0 \in H$ , and a source term  $f \in L^2(]0, T[; H)$ . Then the problem*

$$\begin{cases} \frac{d}{dt} \langle u(t), v \rangle_H + a(u(t), v) = \langle f(t), v \rangle_H & \forall v \in V, \ 0 < t < T, \\ u(t=0) = u_0, \end{cases} \quad (8.7)$$

(where the equation of (8.7) holds in the weak sense in  $]0, T[$ ) has a unique solution  $u \in L^2(]0, T[; V) \cap C([0, T]; H)$ . Further, there exists a constant  $C > 0$  (which only depends on  $\Omega$ ) such that

$$\|u\|_{L^2(]0, T[; V)} + \|u\|_{C([0, T]; H)} \leq C (\|u_0\|_H + \|f\|_{L^2(]0, T[; H)}). \quad (8.8)$$

**Remark 8.2.4** The **energy estimate** (8.8) proves that the solution of (8.7) depends continuously on the data, and therefore the parabolic problem (8.7) is well-posed in the sense of Hadamard. •

**Remark 8.2.5** In theorem 8.2.3 we can weaken the coercivity hypothesis on the symmetric bilinear form  $a(u, v)$  (as we have already proposed in exercise 7.3.1). We obtain the same conclusions by assuming only that there exist two positive constants  $\nu > 0$  and  $\eta > 0$  such that

$$a(v, v) + \eta \|v\|_H^2 \geq \nu \|v\|_V^2 \quad \text{for all } v \in V.$$

In effect, if we make a change of the unknown function  $u(t) = e^{\eta t} w(t)$ , we see that (8.7) is equivalent to

$$\begin{cases} \frac{d}{dt} \langle w(t), v \rangle_H + a(w(t), v) + \eta \langle w(t), v \rangle_H = \langle f(t), v \rangle_H & \forall v \in V, \ 0 < t < T, \\ w(t=0) = u_0, \end{cases}$$

where the bilinear form  $a(w, v) + \eta \langle w, v \rangle_H$  is coercive over  $V$ . This weaker hypothesis is useful, for example, in the solution of exercise 8.2.4. •

**Proof.** The proof is divided into two steps. In the first step, by assuming the existence of a solution  $u$ , we obtain an explicit formula for  $u$  in the form of a series obtained by spectral decomposition of the spaces  $H$  and  $V$ . In particular, this formula proves the uniqueness of the solution. In the second step, we prove that this series converges in the spaces  $L^2(]0, T[; V)$  and  $C([0, T]; H)$ , and that the sum is a solution of (8.7).

**Step 1.** Let us assume that  $u \in L^2(]0, T[; V) \cap C([0, T]; H)$  is a solution of (8.7). The hypotheses allow us to apply theorem 7.3.2 to the solution of the problem to

the eigenvalues associated with the symmetric bilinear form  $a(u, v)$ . Consequently, there exists a Hilbertian basis  $(u_k)_{k \geq 1}$  of  $H$  composed of eigenvectors of (7.12)

$$u_k \in V \quad \text{and} \quad a(u_k, v) = \lambda_k \langle u_k, v \rangle_H \quad \forall v \in V.$$

We define

$$\alpha_k(t) = \langle u(t), u_k \rangle_H, \quad \alpha_k^0 = \langle u_0, u_k \rangle_H, \quad \beta_k(t) = \langle f(t), u_k \rangle_H.$$

Since  $u \in L^2(]0, T[; V) \cap C([0, T]; H)$  and  $f \in L^2(]0, T[; H)$ , we deduce that  $\alpha_k(t) \in C([0, T])$  and  $\beta_k(t) \in L^2(]0, T[)$ . As  $(u_k)_{k \geq 1}$  is a Hilbertian basis of  $H$ , we have

$$u(t) = \sum_{k=1}^{+\infty} \alpha_k(t) u_k,$$

and choosing  $v = u_k$  in (8.7) we obtain

$$\begin{cases} \frac{d\alpha_k}{dt} + \lambda_k \alpha_k = \beta_k & \text{in } ]0, T[ \\ \alpha_k(t=0) = \alpha_k^0. \end{cases} \quad (8.9)$$

We immediately verify that the unique solution of (8.9) is

$$\alpha_k(t) = \alpha_k^0 e^{-\lambda_k t} + \int_0^t \beta_k(s) e^{-\lambda_k(t-s)} ds \quad \text{for } t > 0,$$

which gives an explicit formula for the solution  $u$  (which is therefore unique).

**Step 2.** We shall prove that the series

$$\sum_{j=1}^{+\infty} \left( \alpha_j^0 e^{-\lambda_j t} + \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right) u_j \quad (8.10)$$

converges in  $L^2(]0, T[; V) \cap C([0, T]; H)$  and that its sum, denoted  $u(t)$  is a solution of (8.7). Let us consider the partial sum of order  $k$  of this series

$$w^k(t) = \sum_{j=1}^k \left( \alpha_j^0 e^{-\lambda_j t} + \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right) u_j. \quad (8.11)$$

Clearly  $w^k$  belongs to  $C([0, T]; H)$  since each  $\alpha_j(t)$  is continuous. Let us show that the sequence  $w^k$  is Cauchy in  $C([0, T]; H)$ . For  $l > k$ , by using the orthonormality of

the eigenfunctions  $u_j$ , we have

$$\begin{aligned}
 \|w^l(t) - w^k(t)\|_H &\leq \left\| \sum_{j=k+1}^l \alpha_j^0 e^{-\lambda_j t} u_j \right\|_H + \left\| \sum_{j=k+1}^l \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds u_j \right\|_H \\
 &\leq \left( \sum_{j=k+1}^l |\alpha_j^0|^2 e^{-2\lambda_j t} \right)^{1/2} + \left( \sum_{j=k+1}^l \left( \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right)^2 \right)^{1/2} \\
 &\leq \left( \sum_{j=k+1}^l |\alpha_j^0|^2 \right)^{1/2} + \left( \sum_{j=k+1}^l \frac{1}{2\lambda_j} \int_0^T |\beta_j(s)|^2 ds \right)^{1/2} \\
 &\leq \left( \sum_{j=k+1}^l |\alpha_j^0|^2 \right)^{1/2} + \frac{1}{\sqrt{2\lambda_1}} \left( \sum_{j=k+1}^l \int_0^T |\beta_j(s)|^2 ds \right)^{1/2},
 \end{aligned}$$

since the sequence of the eigenvalues  $(\lambda_j)$  is increasing and strictly positive. As  $u_0 \in H$  and  $f \in L^2([0, T[; H])$  we have

$$\|u_0\|_H^2 = \sum_{j=1}^{+\infty} |\alpha_j^0|^2 < +\infty, \quad \|f\|_{L^2([0, T[; H])}^2 = \sum_{j=1}^{+\infty} \int_0^T |\beta_j(s)|^2 ds < +\infty,$$

which implies that the sequence  $w^k(t)$  is Cauchy in  $H$ . More precisely, we deduce that the sequence  $w^k$  satisfies

$$\lim_{k, l \rightarrow +\infty} \left( \sup_{0 \leq t \leq T} \|w^l - w^k\|_H \right) = 0,$$

that is to say that it is Cauchy in  $C([0, T]; H)$ .

Let us show that the sequence  $w^k$  is also Cauchy in  $L^2([0, T]; V)$ . We equip  $V$  with the scalar product  $a(u, v)$  (equivalent to the usual scalar product because of the coercivity of  $a$ ). For  $l > k$  we have

$$\begin{aligned}
 \|w^l(t) - w^k(t)\|_V^2 &= a(w^l(t) - w^k(t), w^l(t) - w^k(t)) = \sum_{j=k+1}^l \lambda_j |\alpha_j(t)|^2 \\
 &\leq 2 \sum_{j=k+1}^l \lambda_j |\alpha_j^0|^2 e^{-2\lambda_j t} + 2 \sum_{j=k+1}^l \lambda_j \left( \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right)^2.
 \end{aligned}$$

Now, by application of the Cauchy-Schwarz inequality we have

$$\begin{aligned}
 \left( \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right)^2 &\leq \left( \int_0^t |\beta_j(s)|^2 e^{-\lambda_j(t-s)} ds \right) \left( \int_0^t e^{-\lambda_j(t-s)} ds \right) \\
 &\leq \frac{1}{\lambda_j} \left( \int_0^t |\beta_j(s)|^2 e^{-\lambda_j(t-s)} ds \right).
 \end{aligned}$$



Additionally, by the Fubini theorem

$$\begin{aligned} \int_0^T \left( \int_0^t |\beta_j(s)|^2 e^{-\lambda_j(t-s)} ds \right) dt &= \int_0^T |\beta_j(s)|^2 \left( \int_s^T e^{-\lambda_j(t-s)} dt \right) ds \\ &\leq \frac{1}{\lambda_j} \int_0^T |\beta_j(s)|^2 ds. \end{aligned}$$

Consequently, we deduce that

$$\int_0^T \|w^l(t) - w^k(t)\|_V^2 dt \leq \sum_{j=k+1}^l |\alpha_j^0|^2 + \sum_{j=k+1}^l \frac{2}{\lambda_j} \int_0^T |\beta_j(s)|^2 ds,$$

which implies that the sequence  $w^k$  satisfies

$$\lim_{k,l \rightarrow +\infty} \int_0^T \|w^l(t) - w^k(t)\|_V^2 dt = 0,$$

that is to say it is Cauchy in  $L^2([0, T]; V)$ .

As the two spaces  $C([0, T]; H)$  and  $L^2([0, T]; V)$  are complete, the Cauchy sequence  $w^k$  converges and we can define its limit  $u$

$$\lim_{k \rightarrow +\infty} w^k = u \text{ in } C([0, T]; H) \cap L^2([0, T]; V).$$

In particular, as  $w^k(0)$  converges to  $u_0$  in  $H$ , we deduce the desired initial condition,  $u(0) = u_0$  (which is an equality between functions of  $H$ ). On the other hand, it is clear that  $u(t)$ , as the sum of the series (8.10) satisfies the variational formulation (8.7) for every test function  $v = u_k$ . Since  $(u_k/\sqrt{\lambda_k})$  is a Hilbertian basis of  $V$ ,  $u(t)$  therefore satisfies the variational formulation (8.7) for all  $v \in V$ , that is to say that  $u(t)$  is the solution of (8.7).

To obtain the energy estimate (8.8), it is enough to remark that we have proved the bounds

$$\|w^l(t) - w^k(t)\|_H \leq \|u_0\|_H + \frac{1}{\sqrt{2\lambda_1}} \|f\|_{L^2([0, T]; H)}$$

and

$$\int_0^T \|w^l(t) - w^k(t)\|_V^2 dt \leq \|u_0\|_H^2 + \frac{2}{\lambda_1} \|f\|_{L^2([0, T]; H)}^2.$$

By taking  $k = 0$  and making  $l$  tend to infinity, we immediately obtain the desired estimate.  $\square$

**Remark 8.2.6 (delicate)** For the reader enthused by mathematical rigour, we return to the meaning of the time derivative in the variational formulation (8.7). In light of the spaces in which we look for the solution  $u(t)$ , the function  $t \rightarrow \langle u(t), v \rangle_H$  is not differentiable in the classical sense: it only belongs to  $L^2(0, T)$  and to  $C[0, T]$ . We can, nevertheless, define its

derivative in the weak sense of lemma 4.4.12 (or in the sense of distributions). More precisely,  $\frac{d}{dt}\langle u(t), v \rangle_H$  is defined as an element of  $H^{-1}(0, T)$  (that is, to say a continuous linear form over  $H_0^1(0, T)$ ) by the formula

$$\left\langle \frac{d}{dt} \langle u(t), v \rangle_H, \phi(t) \right\rangle_{H^{-1}, H_0^1(0, T)} = - \int_0^T \langle u(t), v \rangle_H \frac{d\phi}{dt}(t) dt \quad \forall \phi \in H_0^1(0, T).$$

Consequently, to say that the equation (8.7) holds in the weak sense in  $]0, T[$  is equivalent to saying that

$$- \int_0^T \langle u(t), v \rangle_H \frac{d\phi}{dt}(t) dt + \int_0^T a(u(t), v) \phi(t) dt = \int_0^T \langle f(t), v \rangle_H \phi(t) dt$$

for all  $v \in V$  and all  $\phi \in C_c^\infty(]0, T[)$  since  $C_c^\infty(]0, T[)$  is dense in  $H_0^1(0, T)$ . To conclude, let us reassure the reader: if  $u$  is a solution of (8.7), then, by (8.7), the derivative  $\frac{d}{dt}\langle u(t), v \rangle_H$  belongs to  $L^2(0, T)$  and we can therefore say that (8.7) holds almost everywhere in  $]0, T[$ . •

### 8.2.3 Applications

We now apply the abstract result of theorem 8.2.3 to the heat flow equation, and we prove that this variational approach allows us to solve the original partial differential equation.

**Theorem 8.2.7** *Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ . Take a final time  $T > 0$ , initial data  $u_0 \in L^2(\Omega)$ , and a source term  $f \in L^2(]0, T[; L^2(\Omega))$ . Then the heat flow equation*

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{a.e. in } \Omega \times ]0, T[ \\ u = 0 & \text{a.e. on } \partial\Omega \times ]0, T[ \\ u(x, 0) = u_0(x) & \text{a.e. in } \Omega. \end{cases} \quad (8.12)$$

*has a unique solution  $u \in L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ . Further, there exists a constant  $C > 0$  (which only depends on  $\Omega$ ) such that, for all  $t \in [0, T]$ ,*

$$\int_{\Omega} u(x, t)^2 dx + \int_0^t \int_{\Omega} |\nabla u(x, s)|^2 dx ds \leq C \left( \int_{\Omega} u_0(x)^2 dx + \int_0^t \int_{\Omega} f(x, s)^2 dx ds \right). \quad (8.13)$$

**Proof.** We apply theorem 8.2.3 to the variational formulation (8.6) of the heat flow equation (8.12): its hypotheses are easily verified with  $H = L^2(\Omega)$  and  $V = H_0^1(\Omega)$  (in particular, as  $\Omega$  is bounded the Rellich theorem 4.3.21 confirms that the injection from  $H$  into  $V$  is compact). It remains to show that the unique solution  $u \in L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$  of this variational formulation is also a solution of (8.12). First, the Dirichlet boundary condition is recovered by application of the trace theorem 4.3.13 to  $u(t) \in H_0^1(\Omega)$  for almost all  $t \in ]0, T[$ , and the initial condition is justified by the continuity of  $u(t)$  at  $t = 0$  (as a function with values in  $L^2(\Omega)$ ).

If the solution  $u$  is sufficiently regular (for example, if  $\frac{\partial u}{\partial t}$  and  $\Delta u$  belong to  $L^2(]0, T[ \times \Omega)$ , which is true from proposition 8.4.6), by integration by parts, the variational formulation (8.6) is equivalent to

$$\int_{\Omega} \left( \frac{\partial u}{\partial t} - \Delta u - f \right) v \, dx = 0, \quad (8.14)$$

for every function  $v(x) \in H_0^1(\Omega)$  and almost all time  $t \in ]0, T[$ . Consequently, we deduce from (8.14) that

$$\frac{\partial u}{\partial t} - \Delta u - f = 0 \quad \text{a.e. in } ]0, T[ \times \Omega.$$

If the solution  $u$  is not more regular than  $u \in L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ , we still obtain this equality but the justification is slightly more delicate. In accordance with remark 8.2.6 the precise meaning of (8.6) is

$$-\int_0^T \int_{\Omega} u v \frac{d\phi}{dt} \, dx \, dt + \int_0^T \int_{\Omega} \nabla u \cdot \nabla v \phi \, dx \, dt = \int_0^T \int_{\Omega} f v \phi \, dx \, dt \quad (8.15)$$

for every function  $v(x) \in C_c^1(\Omega)$  and  $\phi(t) \in C_c^1(]0, T[)$ . A classical result from analysis tells us that the set of linear combinations of products of such functions  $v(x)\phi(t)$  is dense in  $C_c^1(]0, T[ \times \Omega)$ . Denote by  $\sigma = (u, -\nabla u)$  the vector valued function in  $\mathbb{R}^{N+1}$  whose divergence in ‘space-time’ is  $\frac{\partial u}{\partial t} - \Delta u$ . The identity (8.15) tells us that this divergence has a weak meaning (see definition 4.2.6) and is equal to the function  $f$  which belongs to  $L^2(]0, T[; L^2(\Omega))$ , from where we have the equality almost everywhere in  $]0, T[ \times \Omega$ . We must, however, note well that we have shown that the difference  $\frac{\partial u}{\partial t} - \Delta u$  belongs to  $L^2(]0, T[; L^2(\Omega))$ , but not each term individually.  $\square$

**Remark 8.2.8** The **energy estimate** (8.13) shows that the norm of the solution in the energy space is controlled by the norm of the data. It should be noted that this norm does not always correspond to the ‘true’ physical energy (in the case of heat flow the thermal energy is proportional to  $\int_{\Omega} u(t, x) \, dx$ ). The inequality (8.13) has been obtained as a consequence of (8.8), which hides its origin and its physical interpretation. The following exercises allow us to obtain equation (8.13) directly starting from the heat flow equation (8.12) by using an **energy equality** which does nothing but express a physical balance. In particular, these estimates or energy equalities justify the choice of the space  $L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$  in which to look for the solutions since this is precisely **the energy space**, that is, to say the space of minimum regularity in which the energy equalities have a meaning.  $\bullet$

**Exercise 8.2.1** We assume that the hypotheses of theorem 8.2.7 are verified.

1. By assuming that the solution  $u$  of (8.12) is regular enough in  $]0, T[ \times \Omega$ , show that, for all  $t \in [0, T]$ , we have the following energy equality

$$\begin{aligned}
& \frac{1}{2} \int_{\Omega} u(x, t)^2 dx + \int_0^t \int_{\Omega} |\nabla u(x, s)|^2 dx ds \\
&= \frac{1}{2} \int_{\Omega} u_0(x)^2 dx + \int_0^t \int_{\Omega} f(x, s) u(x, s) dx ds.
\end{aligned} \tag{8.16}$$

2. Prove the following property, called 'Gronwall's lemma': if  $z$  is a continuous function of  $[0, T]$  in  $\mathbb{R}^+$  such that

$$z(t) \leq a + b \int_0^t z(s) ds \quad \forall t \in [0, T],$$

where  $a, b$  are two nonnegative constants, then

$$z(t) \leq ae^{bt} \quad \forall t \in [0, T].$$

3. By applying Gronwall's lemma with  $z(t) = \frac{1}{2} \int_{\Omega} u(x, t)^2 dx$ , deduce from (8.16) that, for all  $t \in [0, T]$ ,

$$\begin{aligned}
& \frac{1}{2} \int_{\Omega} u(x, t)^2 dx + \int_0^t \int_{\Omega} |\nabla u(x, s)|^2 dx ds \\
& \leq \frac{e^t}{2} \left( \int_{\Omega} u_0(x)^2 dx + \int_0^T \int_{\Omega} f(x, s)^2 dx ds \right).
\end{aligned} \tag{8.17}$$

**Exercise 8.2.2** In the light of (8.13), where the constant  $C$  is independent of  $T$ , we see that the term  $e^t$  is certainly not optimal in the bound (8.17). This estimate can be improved by reasoning in the following way, with a variant of Gronwall's lemma.

1. Let  $a \in \mathbb{R}^+$  and  $g \in L^2([0, T])$  be such that  $g \geq 0$ . Show that, if  $z(t)$  is continuous from  $[0, T]$  into  $\mathbb{R}^+$  and satisfies

$$z(t) \leq a + 2 \int_0^t g(s) \sqrt{z(s)} ds \quad \forall t \in [0, T],$$

then

$$z(t) \leq \left( \sqrt{a} + \int_0^t g(s) ds \right)^2 \quad \forall t \in [0, T].$$

2. Deduce from (8.16) that, for all  $t \in [0, T]$ ,

$$\begin{aligned}
& \int_{\Omega} u(x, t)^2 dx + 2 \int_0^t \int_{\Omega} |\nabla u(x, s)|^2 dx ds \\
& \leq \left( \left( \int_{\Omega} u_0(x)^2 dx \right)^{1/2} + \int_0^t ds \left( \int_{\Omega} f(x, s)^2 dx \right)^{1/2} \right)^2.
\end{aligned} \tag{8.18}$$

The energy equality (8.16) is not the only one possible for the heat flow equation as the following exercise shows.

**Exercise 8.2.3** We assume that the hypotheses of theorem 8.2.7 are verified, that  $u_0 \in H_0^1(\Omega)$ , and that the solution  $u$  of (8.12) is sufficiently regular in  $]0, T[ \times \Omega$ . Show that, for all  $t \in [0, T]$ , we have the following energy equality

$$\begin{aligned} & \frac{1}{2} \int_{\Omega} |\nabla u(x, t)|^2 dx + \int_0^t \int_{\Omega} \left| \frac{\partial u}{\partial t}(x, s) \right|^2 dx ds \\ &= \frac{1}{2} \int_{\Omega} |\nabla u_0(x)|^2 dx + \int_0^t \int_{\Omega} f(x, s) \frac{\partial u}{\partial t}(x, s) dx ds. \end{aligned} \quad (8.19)$$

Of course, the existence theorem 8.2.7 generalizes easily to the case of other boundary conditions or of a general elliptic operator as the following exercises show.

**Exercise 8.2.4** Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ . Take a final time  $T > 0$ , initial data  $u_0 \in L^2(\Omega)$ , and a source term  $f \in L^2(]0, T[; L^2(\Omega))$ . With the help of remark 8.2.5 show that the heat flow equation with Neumann boundary condition

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{in } \Omega \times ]0, T[ \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega \times ]0, T[ \\ u(x, 0) = u_0(x) & \text{in } \Omega \end{cases} \quad (8.20)$$

has a unique solution  $u \in L^2(]0, T[; H^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ .

**Exercise 8.2.5** Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ . Let  $A(x)$  be a function of  $\Omega$  in the set of the real symmetric matrices such that there exist two constants  $\beta \geq \alpha > 0$  satisfying

$$\beta |\xi|^2 \geq A(x) \xi \cdot \xi \geq \alpha |\xi|^2 \quad \forall \xi \in \mathbb{R}^N, \quad \text{a.e. } x \in \Omega.$$

Take a final time  $T > 0$ , initial data  $u_0 \in L^2(\Omega)$ , and a source term  $f \in L^2(]0, T[; L^2(\Omega))$ . Show that the boundary value problem

$$\begin{cases} \frac{\partial u}{\partial t} - \operatorname{div}(A(x) \nabla u) = f & \text{in } \Omega \times ]0, T[ \\ u = 0 & \text{on } \partial\Omega \times ]0, T[ \\ u(x, 0) = u_0(x) & \text{in } \Omega, \end{cases}$$

has a unique solution  $u \in L^2(]0, T[; H^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ .

We can extend theorem 8.2.7 to the time-dependent Stokes equations.

**Theorem 8.2.9** *Let  $\Omega$  be a regular open bounded connected set of  $\mathbb{R}^N$ . Take a final time  $T > 0$ , initial data  $u_0 \in L^2(\Omega)^N$  such that  $\operatorname{div} u_0 = 0$  in  $\Omega$ , and a source term  $f \in L^2([0, T[; L^2(\Omega))^N$ . Then the time-dependent Stokes equations*

$$\begin{cases} \frac{\partial u}{\partial t} + \nabla p - \mu \Delta u = f & \text{in } \Omega \times ]0, T[ \\ \operatorname{div} u = 0 & \text{in } \Omega \times ]0, T[ \\ u = 0 & \text{on } \partial\Omega \times ]0, T[ \\ u(x, t = 0) = u_0(x) & \text{in } \Omega \end{cases} \quad (8.21)$$

have a unique solution  $u \in L^2([0, T[; H_0^1(\Omega))^N \cap C([0, T]; L^2(\Omega))^N$ .

**Proof.** To obtain a variational formulation of (8.21) we combine the arguments of Section 8.2.1 and of the proof of theorem 5.3.8. We introduce the Hilbert spaces

$$V = \{v \in H_0^1(\Omega)^N \text{ such that } \operatorname{div} v = 0 \text{ a.e. in } \Omega\},$$

and

$$H = \{v \in L^2(\Omega)^N \text{ such that } \operatorname{div} v = 0 \text{ a.e. in } \Omega\},$$

where  $H$  is a closed subspace of  $H(\operatorname{div})$  that we can also define as the adherence of  $V$  in  $L^2(\Omega)^N$  (see Section 4.4.2). We obtain the following variational formulation

$$\begin{cases} \frac{d}{dt} \int_{\Omega} u(t) \cdot v \, dx + \mu \int_{\Omega} \nabla u(t) \cdot \nabla v \, dx = \int_{\Omega} f(t) \cdot v \, dx & \forall v \in V, \quad 0 < t < T, \\ u(t = 0) = u_0, \end{cases} \quad (8.22)$$

where the equation in (8.22) holds in the weak sense in  $]0, T[$ . We apply theorem 8.2.3 to this variational formulation (8.22) (its hypotheses are easily verified) and we obtain the existence and uniqueness of its solution  $u \in L^2([0, T[; H_0^1(\Omega))^N \cap C([0, T]; L^2(\Omega))^N$ .

All the difficulty lies in the proof that this solution of (8.22) is also a solution of (8.21). The Dirichlet boundary condition is recovered by application of the trace theorem 4.3.13 to  $u(t) \in H_0^1(\Omega)^N$  for almost all  $t \in ]0, T[$ , and the initial condition is justified by the continuity of  $u(t)$  at  $t = 0$  since  $u_0 \in H$ .

To recover the equation, we proceed as in the proof of theorem 8.2.7. If the solution  $u$  is sufficiently regular, we obtain

$$\int_{\Omega} \left( \frac{\partial u}{\partial t} - \mu \Delta u - f \right) \cdot v \, dx = 0 \quad (8.23)$$

for almost all  $t \in ]0, T[$ , and arbitrary  $v(x) \in C_c^1(\Omega)^N$  such that  $\operatorname{div} v = 0$  in  $\Omega$ . As for the stationary Stokes problem (see Section 5.3.2), we must deduce from (8.23) the existence of a function  $p(t, x)$  such that the equation (8.21) holds. We must then use Rham's theorem 5.3.9 (or at least one of its variants) which confirms that 'the orthogonal complement to the vectors with zero divergence is the set of the gradients'. This analytical point is quite delicate (even more so if the solution  $u$  is not regular) and we shall simply admit the existence of such a pressure  $p$  without even making clear to which space it belongs.  $\square$

**Remark 8.2.10** Let us briefly mention that there exist other approaches than that used here (and that we can describe as a spectral approach) to obtain the existence and uniqueness of solutions of evolution problems. There exists a purely variational theory (see [28]) as well as a semigroup theory (see [6]). These theories are a little

more complicated, but more powerful since in particular they allow us to ignore the hypotheses on the bounded character of the open set and on the symmetry of the bilinear form of the variational formulation. •

## 8.3 Existence and uniqueness in the hyperbolic case

As in the previous section we follow the same path in three steps. First (Section 8.3.1), we establish a variational formulation, second (Section 8.3.2), we prove the existence and uniqueness of the solution of this variational formulation by using a Hilbertian basis of eigenfunctions, third (Section 8.2.3), we show that this variational solution satisfies the boundary value problem.

### 8.3.1 Variational formulation

The idea is to write a variational formulation which resembles a second order **ordinary differential equation**, similar to (7.4). We therefore multiply the wave equation (8.3) by a test function  $v(x)$  which **does not depend on time**  $t$ . Because of the boundary condition we demand that this  $v$  is zero on the boundary of the open set  $\Omega$ . A formal calculation leads to

$$\frac{d^2}{dt^2} \int_{\Omega} u(x, t) v(x) dx + \int_{\Omega} \nabla u(x, t) \cdot \nabla v(x) dx = \int_{\Omega} f(x, t) v(x) dx. \quad (8.24)$$

It is clear that the ‘natural’ space for the test function  $v$  is  $H_0^1(\Omega)$ . We then introduce the scalar product of  $L^2(\Omega)$  and the bilinear form  $a(w, v)$  defined by

$$\langle w, v \rangle_{L^2(\Omega)} = \int_{\Omega} w(x) v(x) dx \quad \text{and} \quad a(w, v) = \int_{\Omega} \nabla w(x) \cdot \nabla v(x) dx.$$

Take a final time  $T > 0$  (possibly equal to  $+\infty$ ), and a source term  $f \in L^2([0, T]; L^2(\Omega))$ . We are also given initial conditions  $u_0 \in H_0^1(\Omega)$  and  $u_1 \in L^2(\Omega)$ . The variational formulation deduced from (8.24) is therefore: find a solution  $u$  in  $C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  such that

$$\begin{cases} \frac{d^2}{dt^2} \langle u(t), v \rangle_{L^2(\Omega)} + a(u(t), v) = \langle f(t), v \rangle_{L^2(\Omega)} & \forall v \in H_0^1(\Omega), \quad 0 < t < T, \\ u(t=0) = u_0, \quad \frac{du}{dt}(t=0) = u_1. \end{cases} \quad (8.25)$$

The initial data have a meaning in (8.25) thanks to the choice of **the energy space**  $C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  for the solution  $u$ . We shall justify this choice a little more by establishing its links with the energy equalities.

Finally, the time derivative in the variational formulation (8.25) must be taken in the weak sense since *a priori* the function  $t \rightarrow \langle u(t), v \rangle_{L^2(\Omega)}$  is only once differentiable in time since it belongs to  $C^1(0, T)$  (see lemma 4.4.12 and remark 8.2.6 for more detail).

### 8.3.2 A general result

To prove the existence and uniqueness of the solution of the variational formulation (8.25), we return again to the general framework of Section 7.3 to ‘diagonalize’ the Laplacian operator and we are reduced to the solution of a family of simple second order ordinary differential equations. Let  $V$  and  $H$  be two Hilbert spaces such that  $V \subset H$  with dense and compact injection (typically  $V = H_0^1(\Omega)$  and  $H = L^2(\Omega)$ ).

**Theorem 8.3.1** *Let  $V$  and  $H$  two Hilbert spaces such that  $V \subset H$  with compact injection and  $V$  is dense in  $H$ . Let  $a(u, v)$  be a symmetric bilinear form which is continuous and coercive in  $V$ . Take a final time  $T > 0$ , initial data  $(u_0, u_1) \in V \times H$ , and a source term  $f \in L^2(]0, T[; H)$ . Then the problem*

$$\begin{cases} \frac{d^2}{dt^2} \langle u(t), v \rangle_H + a(u(t), v) = \langle f(t), v \rangle_H & \forall v \in V, \quad 0 < t < T, \\ u(t=0) = u_0, \quad \frac{du}{dt}(t=0) = u_1, \end{cases} \quad (8.26)$$

(where the equation of (8.26) holds in the weak sense in  $]0, T[$ ) has a unique solution  $u \in C([0, T]; V) \cap C^1([0, T]; H)$ . Further, there exists a constant  $C > 0$  (which only depends on  $\Omega$  and on  $T$ ) such that

$$\|u\|_{C([0, T]; V)} + \|u\|_{C^1([0, T]; H)} \leq C (\|u_0\|_V + \|u_1\|_H + \|f\|_{L^2(]0, T[; H)}). \quad (8.27)$$

**Remark 8.3.2 The energy estimate** (8.27) proves that the solution of (8.26) depends continuously on the data, and therefore that the hyperbolic problem (8.26) is well-posed in the sense of Hadamard. Proposition 8.3.5 will give an important physical interpretation of a particular case of this energy estimate. •

**Remark 8.3.3** As in the parabolic case (see remark 8.2.5), we can weaken the hypothesis of theorem 8.3.1 on the coercivity of the symmetric bilinear form  $a(u, v)$ . We obtain the same conclusions by assuming only that there exist two positive constants  $\nu > 0$  and  $\eta > 0$  such that

$$a(v, v) + \eta \|v\|_H^2 \geq \nu \|v\|_V^2 \quad \text{for all } v \in V.$$

The change of unknown  $u(t) = e^{\sqrt{\eta}t} w(t)$  transforms the equation of (8.26) into

$$\frac{d^2}{dt^2} \langle w(t), v \rangle_H + 2\sqrt{\eta} \frac{d}{dt} \langle w(t), v \rangle_H + a(w(t), v) + \eta \langle w(t), v \rangle_H = \langle f(t), v \rangle_H, \quad (8.28)$$

where the bilinear form  $a(w, v) + \eta \langle w, v \rangle_H$  is coercive over  $V$ . The equation (8.28) is a damped wave equation (see remark 8.1.1). It is then sufficient to generalize theorem 8.3.1 to such equations (which is easy even though we do not do it here). •



**Proof.** The proof is very similar to that of theorem 8.2.3, so we do not include as much detail. In the first step, we show that every solution  $u$  is a series of eigenfunctions. In the second step, we prove the convergence of this series in the spaces  $C([0, T]; V)$  and  $C^1([0, T]; H)$ .

**Step 1.** Let us assume that  $u \in C([0, T]; V) \cap C^1([0, T]; H)$  is a solution of (8.26). We introduce the Hilbertian basis  $(u_k)_{k \geq 1}$  of  $H$  composed of the eigenfunctions of the variational formulation (7.12) which satisfy

$$u_k \in V \quad \text{and} \quad a(u_k, v) = \lambda_k \langle u_k, v \rangle_H \quad \forall v \in V.$$

We write  $u(t) = \sum_{k=1}^{+\infty} \alpha_k(t) u_k$  with  $\alpha_k(t) = \langle u(t), u_k \rangle_H$ . By choosing  $v = u_k$  in (8.26), and denoting by  $\beta_k(t) = \langle f(t), u_k \rangle_H$ ,  $\alpha_k^0 = \langle u_0, u_k \rangle_H$ , and  $\alpha_k^1 = \langle u_1, u_k \rangle_H$ , we obtain

$$\begin{cases} \frac{d^2 \alpha_k}{dt^2} + \lambda_k \alpha_k = \beta_k & \text{in } ]0, T[ \\ \alpha_k(t=0) = \alpha_k^0, \quad \frac{d\alpha_k}{dt}(t=0) = \alpha_k^1. \end{cases} \quad (8.29)$$

(Note the possible confusion in the notation: the initial data  $u_1$  has nothing to do with the eigenfunction  $u_k$  for  $k = 1$ .) Setting  $\omega_k = \sqrt{\lambda_k}$ , the unique solution of (8.29) is

$$\alpha_k(t) = \alpha_k^0 \cos(\omega_k t) + \frac{\alpha_k^1}{\omega_k} \sin(\omega_k t) + \frac{1}{\omega_k} \int_0^t \beta_k(s) \sin(\omega_k(t-s)) ds, \quad (8.30)$$

which gives an explicit formula for the solution  $u$  (which is therefore unique).

**Step 2.** To prove that the series

$$\sum_{j=1}^{+\infty} \left( \alpha_j^0 \cos(\omega_j t) + \frac{\alpha_j^1}{\omega_j} \sin(\omega_j t) + \frac{1}{\omega_j} \int_0^t \beta_j(s) \sin(\omega_j(t-s)) ds \right) u_j \quad (8.31)$$

converges in  $C([0, T]; V) \cap C^1([0, T]; H)$ , we shall show that the sequence  $w^k = \sum_{j=1}^k \alpha_j(t) u_j$  of the partial sums of this series is Cauchy. In  $V$  we consider the scalar product  $a(u, v)$  for which the family  $(u_j)$  is orthogonal. By the orthogonality of  $(u_j)$  in  $H$  and in  $V$  (see theorem 7.3.2), we obtain, for  $l > k$ , and for all time  $t$ ,

$$a(w^l - w^k, w^l - w^k) + \left\| \frac{d}{dt} (w^l - w^k) \right\|_H^2 = \sum_{j=k+1}^l \left( \lambda_j |\alpha_j(t)|^2 + \left| \frac{d\alpha_j}{dt}(t) \right|^2 \right).$$

Now, multiplying (8.29) by  $d\alpha_k/dt$  and integrating in time, we obtain

$$\left| \frac{d\alpha_j}{dt}(t) \right|^2 + \lambda_j |\alpha_j(t)|^2 = |\alpha_j^1|^2 + \lambda_j |\alpha_j^0|^2 + 2 \int_0^t \beta_j(s) \frac{d\alpha_j}{dt}(s) ds.$$

From formula (8.30) we infer that

$$\left| \frac{d\alpha_j}{dt}(t) \right| \leq \omega_j |\alpha_j^0| + |\alpha_j^1| + \int_0^t |\beta_j(s)| ds.$$

Combining these two results we deduce

$$\left| \frac{d\alpha_j}{dt}(t) \right|^2 + \lambda_j |\alpha_j(t)|^2 \leq 2 |\alpha_j^1|^2 + 2\lambda_j |\alpha_j^0|^2 + 2t \int_0^t |\beta_j(s)|^2 ds. \quad (8.32)$$

As  $u_0 \in V$ ,  $u_1 \in H$  and  $f \in L^2([0, T]; H)$ , we have

$$\|u_0\|_V^2 = a(u_0, u_0) = \sum_{j=1}^{+\infty} \lambda_j |\alpha_j^0|^2 < +\infty, \quad \|u_1\|_H^2 = \sum_{j=1}^{+\infty} |\alpha_j^1|^2 < +\infty,$$

$$\|f\|_{L^2([0, T]; H)}^2 = \sum_{j=1}^{+\infty} \int_0^T |\beta_j(s)|^2 ds < +\infty,$$

which implies that the series, whose general term is the right-hand side of (8.32), is convergent, that is, to say that the sequence  $w^k$  satisfies

$$\lim_{k, l \rightarrow +\infty} \max_{0 \leq t \leq T} \left( \|w^l(t) - w^k(t)\|_V^2 + \left\| \frac{d}{dt}(w^l(t) - w^k(t)) \right\|_H^2 \right) = 0,$$

in other words, it is Cauchy in  $C^1([0, T]; H)$  and in  $C([0, T]; V)$ . Since these spaces are complete, the Cauchy sequence  $w^k$  converges and we can define its limit  $u$ . In particular, as  $(w^k(0), \frac{dw^k}{dt}(0))$  converges to  $(u_0, u_1)$  in  $V \times H$ , we obtain the desired initial conditions. On the other hand, it is clear that  $u(t)$ , as the sum of the series (8.31) satisfies the variational formulation (8.26) for each test function  $v = u_k$ . As  $(u_k/\sqrt{\lambda_k})$  is a Hilbertian basis of  $V$ ,  $u(t)$  therefore satisfies the variational formulation (8.26) for all  $v \in V$ , that is, to say that  $u(t)$  is the solution we seek of (8.26). Additionally, we have in fact shown that

$$a(w^l - w^k, w^l - w^k) + \left\| \frac{d}{dt}(w^l - w^k) \right\|_H^2 \leq C \left( \|u_0\|_V^2 + \|u_1\|_H^2 + T \|f\|_{L^2([0, T]; H)}^2 \right),$$

and the energy estimate (8.27) is then easily obtained by taking  $k = 0$  and letting  $l$  tend to infinity.  $\square$

### 8.3.3 Applications

We now apply the abstract result of theorem 8.3.1 to the wave equation, and we prove that this variational approach allows us to solve the original partial differential equation.

**Theorem 8.3.4** *Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ , and take a final time  $T > 0$ . We consider initial data  $(u_0, u_1) \in H_0^1(\Omega) \times L^2(\Omega)$  and a source term  $f \in L^2([0, T]; L^2(\Omega))$ . Then the wave equation*

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta u = f & \text{a.e. in } \Omega \times ]0, T[ \\ u = 0 & \text{a.e. on } \partial\Omega \times ]0, T[ \\ u(x, 0) = u_0(x) & \text{a.e. in } \Omega \\ \frac{\partial u}{\partial t}(x, 0) = u_1(x) & \text{a.e. in } \Omega. \end{cases} \quad (8.33)$$

*has a unique solution  $u \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ . Further, there exists a constant  $C > 0$  (which only depends on  $\Omega$  and on  $T$ ) such that, for all  $t \in [0, T]$ ,*

$$\begin{aligned} & \int_{\Omega} \left( \left| \frac{\partial u}{\partial t}(x, t) \right|^2 + |\nabla u(x, t)|^2 \right) dx \\ & \leq C \left( \int_{\Omega} (|u_1(x)|^2 + |\nabla u_0(x)|^2) dx + \int_0^t \int_{\Omega} |f(x, s)|^2 dx ds \right). \end{aligned} \quad (8.34)$$

**Proof.** We apply theorem 8.3.1 to the variational formulation (8.25) of the wave equation obtained in Section 8.3.1 (its hypotheses are easily verified with  $H = L^2(\Omega)$  and  $V = H_0^1(\Omega)$ ). It remains to show that the unique solution  $u \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  of this variational formulation is also a solution of (8.33). First of all, the Dirichlet boundary conditions are recovered by application of the trace theorem 4.3.13 to  $u(t) \in H_0^1(\Omega)$  for all  $t \in [0, T]$ , and the initial condition is justified by the continuity of  $u(t)$  at  $t = 0$  as a function with values in  $H_0^1(\Omega)$  and of  $du/dt(t)$  at  $t = 0$  as a function with values in  $L^2(\Omega)$ .

If the solution  $u$  is sufficiently regular, by integration by parts the variational formulation (8.25) is equivalent to

$$\int_{\Omega} \left( \frac{\partial^2 u}{\partial t^2} - \Delta u - f \right) v dx = 0,$$

for all  $v(x) \in C_c^1(\Omega)$  and almost all  $t \in ]0, T[$ . We therefore deduce the equation in (8.33). If the solution  $u$  is not more regular than is given by theorem 8.3.1, we still obtain the equation ‘almost everywhere’, by taking the arguments of the proof of theorem 8.2.7 (which we do not detail). We denote by  $\sigma = (\frac{\partial u}{\partial t}, -\nabla u)$  the function with vector values in  $\mathbb{R}^{N+1}$ , and we can show that it has a weak divergence in ‘space-time’ which is exactly  $\frac{\partial^2 u}{\partial t^2} - \Delta u$  which therefore belongs to  $L^2([0, T]; L^2(\Omega))$ .  $\square$

In the absence of forces,  $f = 0$ , we can improve the energy estimate (8.34) and obtain a property of **conservation of total energy** which is very important from the point of view of applications. The total energy is here the sum of two terms: on the one hand the kinetic energy  $|\frac{\partial u}{\partial t}|^2$  and on the other hand the mechanical energy  $|\nabla u|^2$ .

**Proposition 8.3.5** *We use the hypotheses of theorem 8.3.4 with  $f = 0$ . The solution of the wave equation (8.33) satisfies, for all  $t \in [0, T]$ , the conservation of energy equation*

$$\int_{\Omega} \left( \left| \frac{\partial u}{\partial t}(x, t) \right|^2 + |\nabla u(x, t)|^2 \right) dx = \int_{\Omega} (|u_1(x)|^2 + |\nabla u_0(x)|^2) dx. \quad (8.35)$$

**Proof.** By taking the proof of theorem 8.3.1 with  $f = 0$ , that is to say  $\beta_k = 0$ , we deduce directly from (8.29) that the energy of the harmonic oscillator is conserved, that is to say that

$$\left| \frac{d\alpha_j}{dt}(t) \right|^2 + \lambda_j |\alpha_j(t)|^2 = |\alpha_j^1|^2 + \lambda_j |\alpha_j^0|^2,$$

which gives the equality (rather than the inequality)

$$a(w^l - w^k, w^l - w^k) + \left\| \frac{d}{dt}(w^l - w^k) \right\|_H^2 = \sum_{j=k+1}^l |\alpha_j^1|^2 + \lambda_j |\alpha_j^0|^2,$$

and (8.35) is obtained by taking  $k = 0$  and letting  $l$  tend to infinity. If the solution  $u$  is regular, we can prove (8.35) more directly by multiplying the wave equation (8.33) by  $\frac{\partial u}{\partial t}$  and integrating by parts (see exercise 8.3.1).  $\square$

We now return to **the energy estimate** (8.34) in the general case  $f \neq 0$ . The following exercise shows how (8.34) can be obtained directly starting from (8.33) with the help of a similar argument to that of proposition 8.3.5 which establishes an **energy equality** which does nothing but express a physical balance. In particular, these estimates or energy equalities justify the choice of the space  $C([0, T]; H_0^1(\Omega))^N \cap C^1([0, T]; L^2(\Omega))^N$  where we look for the solutions since this is precisely **the energy space** that is to say the space of minimum regularity in which the energy equalities have a meaning.

**Exercise 8.3.1** We assume that the hypotheses of theorem 8.3.4 are verified.

1. By assuming that the solution  $u$  of (8.33) is regular enough in  $]0, T[ \times \Omega$ , show that, for all  $t \in [0, T]$ , we have the following energy equality

$$\begin{aligned} & \int_{\Omega} \left| \frac{\partial u}{\partial t}(x, t) \right|^2 dx + \int_{\Omega} |\nabla u(x, t)|^2 dx \\ &= \int_{\Omega} u_1(x)^2 dx + \int_{\Omega} |\nabla u_0(x)|^2 dx + 2 \int_0^t \int_{\Omega} f(x, s) \frac{\partial u}{\partial t}(x, s) dx ds. \end{aligned}$$

2. Deduce that there exists a constant  $C(T)$  (independent of the data other than  $T$ ) such that

$$\begin{aligned} & \int_{\Omega} \left| \frac{\partial u}{\partial t}(x, t) \right|^2 dx + \int_{\Omega} |\nabla u(x, t)|^2 dx \\ & \leq C(T) \left( \int_{\Omega} u_1(x)^2 dx + \int_{\Omega} |\nabla u_0(x)|^2 dx + \int_0^t \int_{\Omega} f(x, s)^2 dx ds \right). \end{aligned}$$

3. Show that there exists a constant  $C$  (independent of all data including  $T$ ) such that

$$\begin{aligned} & \int_{\Omega} \left| \frac{\partial u}{\partial t}(x, t) \right|^2 dx + \int_{\Omega} |\nabla u(x, t)|^2 dx \\ & \leq C \left( \int_{\Omega} u_1(x)^2 dx + \int_{\Omega} |\nabla u_0(x)|^2 dx + \left( \int_0^t \left( \int_{\Omega} f(x, s)^2 dx \right)^{1/2} ds \right)^2 \right). \end{aligned}$$

Other conserved quantities exist as is shown in the following exercise.

**Exercise 8.3.2** We assume that the hypotheses of theorem 8.3.4 are verified, that the source term is zero,  $f = 0$ , and that the solution  $u$  of (8.33) is regular in  $[0, T] \times \Omega$ . Show that, for every integer  $m \geq 1$ , we have

$$\frac{d}{dt} \int_{\Omega} \left( \left| \frac{\partial^m u}{\partial t^m} \right|^2 + \left| \nabla \frac{\partial^{m-1} u}{\partial t^{m-1}} \right|^2 \right) dx = 0.$$

Of course, existence theorem 8.3.4 generalizes easily to the case of other boundary conditions (for example, Neumann), or to the case of operators other than the Laplacian like

$$\frac{\partial^2 u}{\partial t^2} - \operatorname{div} (A(x) \nabla u) = f.$$

The following is an exercise to generalize this result to the elastodynamic equations.

**Exercise 8.3.3** Let  $\Omega$  be a regular open bounded connected set of  $\mathbb{R}^N$ . Take the initial data  $(u_0, u_1) \in H_0^1(\Omega)^N \times L^2(\Omega)^N$ , and a source term  $f \in L^2([0, T]; L^2(\Omega))^N$ . Show that there exists a unique solution  $u \in C([0, T]; H_0^1(\Omega))^N \cap C^1([0, T]; L^2(\Omega))^N$  of

$$\begin{cases} \rho \frac{\partial^2 u}{\partial t^2} - \operatorname{div} (2\mu e(u) + \lambda \operatorname{tr}(e(u)) \mathbf{I}) = f & \text{in } \Omega \times ]0, T[, \\ u = 0 & \text{on } \partial\Omega \times ]0, T[, \\ u(t = 0) = u_0(x) & \text{in } \Omega, \\ \frac{\partial u}{\partial t}(t = 0) = u_1(x) & \text{in } \Omega. \end{cases} \quad (8.36)$$

(As usual, the Lamé coefficients satisfy  $\mu > 0$  and  $2\mu + N\lambda > 0$ .)

By assuming that the solution  $u$  is sufficiently regular, show that, for all  $t \in [0, T]$ , we have the following energy equality

$$\begin{aligned} & \frac{\rho}{2} \int_{\Omega} \left| \frac{\partial u}{\partial t} \right|^2 dx + \mu \int_{\Omega} |e(u)|^2 dx + \frac{\lambda}{2} \int_{\Omega} (\operatorname{div} u)^2 dx \\ &= \frac{\rho}{2} \int_{\Omega} |u_1|^2 dx + \mu \int_{\Omega} |e(u_0)|^2 dx + \frac{\lambda}{2} \int_{\Omega} (\operatorname{div} u_0)^2 dx + \int_0^t \int_{\Omega} f \cdot \frac{\partial u}{\partial t} dx ds. \end{aligned}$$

Deduce an energy estimate.

**Remark 8.3.6** As for parabolic problems, the ‘spectral’ approach used here to obtain the existence and uniqueness of hyperbolic partial differential equations is not the only one possible. Let us cite the purely variational theory of [28] and also the semigroup theory (see [6]). These theories are a little more complicated, but more powerful since in particular we do not need the hypotheses on the bounded character of the open set  $\Omega$  and on the symmetry of the bilinear form  $a(u, v)$  of the variational formulation. •

## 8.4 Qualitative properties in the parabolic case

We now examine the principal qualitative properties of the solution of the heat flow equation, notably the properties of regularity, asymptotic behaviour for large values of  $t$ , and the maximum principle.

### 8.4.1 Asymptotic behaviour

We study the behaviour of the solution of the heat flow equation for large time, that is to say when  $t$  tends to  $+\infty$ . We shall verify that, agreeing with physical intuition, if the right-hand side  $f(x)$  is independent of time  $t$ , then the solution of the heat flow equation tends asymptotically to the (stationary) solution of the Laplacian. We start by examining the case of the homogeneous heat flow equation.

**Proposition 8.4.1** *Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ . Take  $u_0 \in L^2(\Omega)$  and  $u$  the solution of the problem*

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = 0 & \text{in } ]0, +\infty[ \times \Omega \\ u(x, t) = 0 & \text{on } ]0, +\infty[ \times \partial\Omega \\ u(x, 0) = u_0(x) & \text{in } \Omega. \end{cases} \quad (8.37)$$

*Then,  $u(t)$  converges to zero in  $L^2(\Omega)$  as  $t$  tends to  $+\infty$*

$$\lim_{t \rightarrow +\infty} \|u(t)\|_{L^2(\Omega)} = 0. \quad (8.38)$$

**Proof.** We return to the proof of theorem 8.2.3 in the case  $f = 0$ , that is to say  $\beta_k = 0$ . We easily see that the partial sum satisfies

$$\|w^l(t) - w^k(t)\|_H^2 = \sum_{j=k+1}^l |\alpha_j^0|^2 e^{-2\lambda_j t},$$

with  $H = L^2(\Omega)$ , which leads, by taking  $k = 0$  and  $l = +\infty$ , and by bounding, to

$$\|u(t)\|_H^2 \leq \|u_0\|_H^2 e^{-2\lambda_1 t}$$

which tends to zero as  $t$  tends to infinity as  $\lambda_1 > 0$ .  $\square$

The case of a nonzero right-hand side, independent of time, is a simple exercise that we leave to the reader.

**Exercise 8.4.1** We use the hypotheses of proposition 8.4.1. Let  $f(x) \in L^2(\Omega)$  and  $u(t, x)$  the solution of

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{in } ]0, +\infty[ \times \Omega \\ u(x, t) = 0 & \text{on } ]0, +\infty[ \times \partial\Omega \\ u(x, 0) = u_0(x) & \text{in } \Omega. \end{cases}$$

Let  $v(x) \in H_0^1(\Omega)$  be the solution of

$$\begin{cases} -\Delta v = f & \text{in } \Omega \\ v = 0 & \text{on } \partial\Omega. \end{cases}$$

Show that  $\lim_{t \rightarrow +\infty} \|u(x, t) - v(x)\|_{L^2(\Omega)} = 0$ .

We can in fact specify the conclusion of proposition 8.4.1 as the following exercise shows whose interpretation is the following

$$u(t, x) \approx \left( \int_{\Omega} u_0 u_1 dx \right) e^{-\lambda_1 t} u_1(x) \quad \text{as } t \rightarrow +\infty,$$

where  $u_1(x)$  is the first (normalized) eigenfunction of the Laplacian 7.17. Asymptotically, all the solutions of the homogeneous heat flow equation decrease exponentially in time with the same spatial profile which is given by  $u_1$  (no matter what the initial data).

**Exercise 8.4.2** We use the hypotheses of proposition 8.4.1. Show that there exists a positive constant  $C$  such that

$$\|u(t) - \alpha_1^0 e^{-\lambda_1 t} u_1\|_{L^2(\Omega)} \leq C e^{-\lambda_2 t} \quad \forall t > 1, \quad \text{with } \alpha_1^0 = \int_{\Omega} u_0 u_1 dx, \quad (8.39)$$

where  $\lambda_k$  denotes the  $k$ th eigenvalue of the Laplacian with Dirichlet boundary condition.

### 8.4.2 The maximum principle

For the heat flow equation, the maximum principle takes a form close to that we have stated in theorem 5.2.22 for the Laplacian.

**Proposition 8.4.2** *Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ , and take a final time  $T > 0$ . Take  $u_0 \in L^2(\Omega)$ ,  $f \in L^2([0, T]; L^2(\Omega))$ , and  $u \in C([0, T]; L^2(\Omega)) \cap L^2([0, T]; H_0^1(\Omega))$  the unique solution of (8.12). If  $f \geq 0$  almost everywhere in  $[0, T] \times \Omega$  and  $u_0 \geq 0$  almost everywhere in  $\Omega$ , then  $u \geq 0$  almost everywhere in  $[0, T] \times \Omega$ .*

**Proof.** Take  $u^- = \min(u, 0)$  which belongs to  $L^2([0, T]; H_0^1(\Omega))$  from lemma 5.2.24 and which satisfies, for  $0 < t < T$ ,

$$\int_{\Omega} \nabla u(t) \cdot \nabla u^-(t) dx = \int_{\Omega} |\nabla u^-(t)|^2 dx. \quad (8.40)$$

An argument similar to that which allowed us to prove (8.40) shows that, if  $\frac{\partial u}{\partial t} \in L^2([0, T]; L^2(\Omega))$ , then

$$\int_{\Omega} \frac{\partial u}{\partial t}(t) u^-(t) dx = \frac{1}{2} \frac{d}{dt} \left( \int_{\Omega} |u^-(t)|^2 dx \right). \quad (8.41)$$

We admit that the identity (8.41) remains true even if  $\frac{\partial u}{\partial t}$  does not belong to  $L^2([0, T]; L^2(\Omega))$ . Consequently, by taking  $v = u^-$  in the variational formulation (8.6) of the heat flow equation we obtain

$$\frac{1}{2} \frac{d}{dt} \left( \int_{\Omega} |u^-|^2 dx \right) + \int_{\Omega} |\nabla u^-|^2 dx = \int_{\Omega} f u^- dx,$$

which gives by integration in time

$$\frac{1}{2} \int_{\Omega} |u^-(t)|^2 dx + \int_0^t \int_{\Omega} |\nabla u^-|^2 dx ds = \int_0^t \int_{\Omega} f u^- dx ds + \frac{1}{2} \int_{\Omega} |u^-(0)|^2 dx.$$

As  $u^-(0) = (u_0)^- = 0$  we deduce

$$\frac{1}{2} \int_{\Omega} |u^-(t)|^2 dx + \int_0^t \int_{\Omega} |\nabla u^-|^2 dx ds \leq 0,$$

that is to say that  $u^- = 0$  almost everywhere in  $[0, T] \times \Omega$ .  $\square$

As in the elliptic case, the maximum principle given by proposition 8.4.2 conforms to physical intuition. In the framework of the model of heat flow described in Chapter 1, if the initial temperature  $u_0(x)$  is at every point larger than the value 0 at which we maintain the temperature on the boundary  $\partial\Omega$  and if the source term is positive (corresponding to a heating effect), then it is clear that the temperature is



positive at every point and at every instant. We have only checked that the mathematical model reproduces this intuitive property.

It is good to realize that these results can also be stated in several equivalent forms. We can, for example, compare two solutions of (8.12): if  $u_0 \leq \tilde{u}_0$  in  $\Omega$  and  $f \leq \tilde{f}$  in  $]0, T[ \times \Omega$ , and if  $u$  and  $\tilde{u}$  denote the solutions of (8.12) corresponding to the data  $(u_0, f)$  and  $(\tilde{u}_0, \tilde{f})$  respectively, then we have  $u \leq \tilde{u}$  in  $]0, T[ \times \Omega$ .

The two following exercises illustrate some interesting applications of the maximum principle.

**Exercise 8.4.3** Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ . We denote by  $u_1$  the first eigenfunction of the Laplacian in  $\Omega$  with Dirichlet conditions, and by  $\lambda_1$  the associated eigenvalue. We recall that we can choose  $u_1 > 0$  in  $\Omega$  (see the Krein–Rutman theorem 7.3.10) and we also have  $\partial u_1 / \partial n > 0$  on  $\partial\Omega$ . Take  $f = 0$ ,  $u_0 \in L^2(\Omega)$  and  $u$  the unique solution (assumed regular) of (8.12).

Take  $\epsilon > 0$ . Show that we can find a positive constant  $K$  such that

$$-Ku_1(x) \leq u(x, \epsilon) \leq Ku_1(x) \quad \forall x \in \bar{\Omega}, \quad (8.42)$$

and deduce that there exists a positive constant  $C$  such that

$$\max_{x \in \bar{\Omega}} |u(x, t)| \leq Ce^{-\lambda_1 t} \quad \forall t > \epsilon. \quad (8.43)$$

**Exercise 8.4.4** Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ . Take  $u_0 \in L^\infty(\Omega)$ ,  $f \in L^\infty(\mathbb{R}^+ \times \Omega)$ , and  $u \in C([0, T]; L^2(\Omega)) \cap L^2([0, T]; H_0^1(\Omega))$  the unique solution of (8.12). Show that

$$\|u\|_{L^\infty(\mathbb{R}^+ \times \Omega)} \leq \|u_0\|_{L^\infty(\Omega)} + \frac{D^2}{2N} \|f\|_{L^\infty(\mathbb{R}^+ \times \Omega)}, \quad (8.44)$$

where  $D = \sup_{x, y \in \Omega} |x - y|$  is the diameter of  $\Omega$ . First consider the easy case where  $f \equiv 0$ , then, in the general case, use the function  $\psi \in H_0^1(\Omega)$  such that  $-\Delta\psi = 1$  in  $\Omega$ .

### 8.4.3 Propagation at infinite velocity

We have already mentioned in remark 1.2.9 this surprising property of the heat flow equation: the heat propagates at infinite velocity! This result follows from a **strong maximum principle** that we now state without proof. We verify it more easily when the domain  $\Omega$  is the entire space  $\mathbb{R}^N$  (see Section 8.4.5).

**Proposition 8.4.3** Let  $\Omega$  be a regular open bounded set of class  $\mathcal{C}^2$  of  $\mathbb{R}^N$ . Take a final time  $T > 0$ . With  $u_0 \in L^2(\Omega)$  and  $u$  the unique solution in  $C([0, T]; L^2(\Omega)) \cap L^2([0, T]; H_0^1(\Omega))$  of the problem

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = 0 & \text{in } ]0, T[ \times \Omega \\ u(x, t) = 0 & \text{on } ]0, T[ \times \partial\Omega \\ u(x, 0) = u_0(x) & \text{in } \Omega. \end{cases}$$

We further assume that  $u_0(x) \geq 0$  almost everywhere in  $\Omega$  and that  $u_0$  is not identically zero. Then, for all time  $\epsilon > 0$ , we have

$$u(x, \epsilon) > 0 \quad \forall x \in \Omega. \quad (8.45)$$

Inequality (8.45) is **strict** which is remarkable (we already had a weaker inequality from the maximum principle of proposition 8.4.2). In effect, if  $u_0$  has compact support in  $\Omega$  and if we look at a point  $x \in \Omega$  outside of the support of  $u_0$ , we find that  $u(x, \epsilon) > 0$  although initially  $u_0(x) = 0$ . In other words, in the framework of the modelling of the temperature evolution, even if the point  $x$  is initially cold ( $u_0(x) = 0$ ) and very far from the initial hot part (the support of  $u_0$ ), it becomes immediately hot as for all time  $t = \epsilon$  (even very small) we have  $u(x, \epsilon) > 0$ . Therefore **the heat propagates at infinite velocity** since its effect is immediate even at a great distance! This is clearly a fault of the mathematical model as we know that nothing can propagate more quickly than the velocity of light (in fact it is the Fourier law (1.3) which is wrong). This is a model, qualitatively and quantitatively correct in many regards, as we have shown in several preceding results, which conforms to physical intuition, but this is only an idealized model of reality.

**Remark 8.4.4** The same property of ‘propagation at infinite velocity’ can also be observed for the Stokes equations (8.2). In this framework, we are perhaps less surprised by this paradox if we realize that the incompressibility hypothesis of a fluid implies that the velocity of sound is infinite. In other words, by using the approximation of incompressibility, we implicitly introduced in the model of the possibility of propagation of the information at infinite velocity. •

#### 8.4.4 Regularity and regularizing effect

In the elliptic case we have seen that the regularity of the solution is directly linked to that of the data. In the parabolic case, the situation is different since, if the source term is zero ( $f = 0$ ), there exists a **regularizing effect** on the initial condition: surprisingly, even if the initial data  $u_0$  is not very regular, the solution immediately becomes very regular.

**Proposition 8.4.5** *Let  $\Omega$  be a regular open bounded set of class  $C^\infty$  of  $\mathbb{R}^N$ , and take a final time  $T > 0$ . Take  $u_0 \in L^2(\Omega)$ , and  $u$  the unique solution in  $C([0, T]; L^2(\Omega)) \cap L^2(]0, T[; H_0^1(\Omega))$  of*

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = 0 & \text{in } ]0, T[ \times \Omega \\ u(x, t) = 0 & \text{on } ]0, T[ \times \partial\Omega \\ u(x, 0) = u_0(x) & \text{in } \Omega. \end{cases} \quad (8.46)$$

*Then, for all  $\epsilon > 0$ ,  $u$  is of class  $C^\infty$  in  $x$  and  $t$  in  $\bar{\Omega} \times [\epsilon, T]$ .*

**Outline of proof.** Rather than a rigorous (and quite technical, see exercise 8.4.5) proof we propose a formal calculation which shows the essential idea behind this

regularity result (the proof is easier in the case  $\Omega = \mathbb{R}^N$ , see exercise 8.4.9). For  $k \geq 1$  we denote by  $v = \frac{\partial^k u}{\partial t^k}$  and we (formally) differentiate the heat flow equation (8.46)  $k$  times with respect to time to obtain

$$\begin{cases} \frac{\partial v}{\partial t} - \Delta v = 0 & \text{in } ]0, T[ \times \Omega \\ v(x, t) = 0 & \text{on } ]0, T[ \times \partial\Omega \\ v(x, 0) = \frac{\partial^k u}{\partial t^k}(0, x) & \text{in } \Omega, \end{cases} \quad (8.47)$$

which is again a heat flow equation. If  $\frac{\partial^k u}{\partial t^k}(0, x)$  belongs to  $L^2(\Omega)$ , we apply the existence and uniqueness theorem 8.2.7 to (8.47) which tells us that  $v$  belongs to  $L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ . In particular,  $u$  is regular in time. On the other hand, by equality,  $v = (\Delta)^k u$  belongs to the same space. By elliptic regularity (see theorem 5.2.26) we deduce that  $u$  is regular in space. The more delicate point in order to be able to give a meaning to this formal reasoning is that the initial data of (8.47) is not very regular. It is for this reason that the regularity of  $u$  is only valid for the time  $t > \epsilon > 0$ .  $\square$

In the presence of source terms, the same reasoning as that of the proof of proposition 8.4.5 allows us to recover a more classical regularity result which is not unconnected to the energy equality (8.19) (in effect, the space to which  $u$  will belong is that which gives a meaning to (8.19)).

**Proposition 8.4.6** *Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ , and take a final time  $T > 0$ . For a source term  $f \in L^2(]0, T[; L^2(\Omega))$  and regular initial data  $u_0 \in H_0^1(\Omega)$ , we consider the unique solution  $u \in L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$  of the heat flow equation (8.12). Then, this solution is more regular in the sense that  $\frac{\partial u}{\partial t} \in L^2(]0, T[; L^2(\Omega))$  and  $u \in L^2(]0, T[; H^2(\Omega)) \cap C([0, T]; H_0^1(\Omega))$ .*

**Remark 8.4.7** We can of course ‘increase’ regularity and find that the solution  $u$  of the heat flow equation (8.12) is as regular as we want, provided that the data  $u_0$  and  $f$  are also regular (see [6]). However, if we want the solution  $u$  to be regular from the initial instant, the data  $u_0$  and  $f$  must satisfy compatibility conditions. Thus, in proposition 8.4.6 it is necessary that the initial condition  $u_0$  satisfies the Dirichlet boundary condition (which has not been necessary for the existence of a solution in theorem 8.2.7). The other compatibility conditions are obtained by remarking that the successive derivatives of  $u$  with respect to time  $t$  are also solutions of the heat flow equation with Dirichlet boundary conditions. For example, the initial condition for the first derivative is  $\frac{\partial u}{\partial t}(0) = f(0) + \Delta u_0$ . So that  $\frac{\partial u}{\partial t}$  is regular, this initial data must therefore satisfy the Dirichlet boundary condition  $f(0) + \Delta u_0 = 0$  on  $\partial\Omega$ , which is a compatibility condition between  $u_0$  and  $f$ .  $\bullet$

**Exercise 8.4.5 (difficult)** Prove proposition 8.4.5 rigorously. For this we shall introduce, for every integer  $m \geq 0$ , the space

$$W^{2m}(\Omega) = \{v \in H^{2m}(\Omega), v = \Delta v = \dots \Delta^{m-1}v = 0 \text{ on } \partial\Omega\}, \quad (8.48)$$

which we equip with the norm  $\|v\|_{W^{2m}(\Omega)}^2 = \int_{\Omega} |(\Delta)^m v|^2 dx$ . Show it is equivalent to the norm of  $H^{2m}(\Omega)$ . Revisit the proof of theorem 8.2.3 by showing that the sequence  $(w_k)$  of the partial sums is Cauchy in  $C^\ell([ \epsilon, T ], W^{2m}(\Omega))$ .

### 8.4.5 Heat equation in the entire space

To finish this section, we briefly indicate how to solve the heat flow equation posed in the whole space  $\mathbb{R}^N$ . Let us recall that the spectral approach followed in this chapter is limited to the case of open bounded sets (this limitation is in fact artificial and absolutely unnecessary to establish the existence and uniqueness of the solution of a parabolic equation). Let us consider the homogeneous heat flow equation in the entire space  $\mathbb{R}^N$ , equipped with initial data

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = 0 & \text{in } ]0, +\infty[ \times \mathbb{R}^N \\ u(x, 0) = u_0(x) & \text{in } \mathbb{R}^N. \end{cases} \quad (8.49)$$

The following classical result shows that the solution of the problem (8.49) is given explicitly as the convolution of the initial data  $u_0$  with a Gaussian whose standard deviation grows like  $\sqrt{t}$ .

**Theorem 8.4.8** *We assume that  $u_0 \in L^2(\mathbb{R}^N)$ . Then problem (8.49) has a unique solution  $u \in C(\mathbb{R}^+, L^2(\mathbb{R}^N)) \cap C^1(\mathbb{R}_*^+, L^2(\mathbb{R}^N))$ , given by*

$$u(x, t) = \frac{1}{(4\pi t)^{N/2}} \int_{\mathbb{R}^N} u_0(y) e^{-|x-y|^2/4t} dy. \quad (8.50)$$

**Proof.** For  $t \geq 0$ , we introduce the Fourier transform of  $u(t)$  (see [4]), that is to say of the function  $x \mapsto u(x, t)$ , defined by

$$\hat{u}(k, t) = \frac{1}{(2\pi)^{N/2}} \int_{\mathbb{R}^N} u(x, t) e^{ik \cdot x} dx,$$

for  $k \in \mathbb{R}^N$ . If  $u \in C(\mathbb{R}^+, L^2(\mathbb{R}^N)) \cap C^1(\mathbb{R}_*^+, L^2(\mathbb{R}^N))$  satisfies (8.49), we can apply the Fourier transform to the two equations (8.49) to obtain

$$\begin{cases} \hat{u} \in C(\mathbb{R}^+, L^2(\mathbb{R}^N)) \cap C^1(\mathbb{R}_*^+, L^2(\mathbb{R}^N)), \\ \frac{\partial \hat{u}}{\partial t} + |k|^2 \hat{u} = 0 & \text{for } k \in \mathbb{R}^N, t > 0, \\ \hat{u}(k, 0) = \hat{u}_0(k) & \text{for } k \in \mathbb{R}^N, \end{cases} \quad (8.51)$$

where  $\hat{u}_0(k) = (1/(2\pi)^{N/2}) \int_{\mathbb{R}^N} u_0(x) e^{ik \cdot x} dx$  is the Fourier transform of  $u_0$ . The system (8.51) is easily solved since we have a differential equation for each value of  $k$ . We obtain

$$\hat{u}(k, t) = \hat{u}_0(k) e^{-|k|^2 t} \quad \text{for } (k, t) \in \mathbb{R}^N \times \mathbb{R}^+,$$

and it is easy to deduce (8.50) by inverse Fourier transform (since this transformation changes a convolution product to a simple product).  $\square$

**Remark 8.4.9** The use of the Fourier transform allows us to ‘diagonalize’ the heat flow equation (8.49) and to reduce the problem to the solution of a simple ordinary differential equation (8.51). This method is therefore very similar, in spirit, to the spectral approach used before and which also relies on a diagonalization argument. In others terms,  $|k|^2$  and  $e^{ik \cdot x}$  are interpreted as eigenvalues and eigenfunctions of the Laplacian in  $\mathbb{R}^N$ . Let us remark that the larger the Fourier mode  $|k|$ , the faster is the exponential decrease in time of  $\hat{u}(k, t)$ : this more rapid damping for small wavelengths ( $k$  large) is linked to the regularizing effect of the heat equation see exercise 8.4.9 below).  $\bullet$

**Remark 8.4.10** The result (8.50) can be interpreted in terms of a Green's function. By setting  $G(x, t) = (1/(2\pi t)^{N/2})e^{-|x|^2/4t}$ , we write (8.50) in the form  $u(t) = G(t) * u_0$ , that is to say

$$u(x, t) = \int_{\mathbb{R}^N} G(x - y, t) u_0(y) dy.$$

The Green's function is the elementary solution of the heat flow equation in  $\mathbb{R}^N \times \mathbb{R}_*^+$ . This means that we can verify that

$$\begin{cases} \frac{\partial G}{\partial t} - \Delta G = 0 & \text{in } ]0, +\infty[ \times \mathbb{R}^N \\ G(x, 0) = \delta_0(x) & \text{in } \mathbb{R}^N. \end{cases}$$

in the sense of distributions, where  $\delta_0$  is the Dirac mass at the origin. This point of view can also be developed in a bounded domain and leads to a method of solution for parabolic equations different from the 'spectral' approach followed here. •

The following exercise allows us to solve the nonhomogeneous heat flow equation.

**Exercise 8.4.6** For  $u_0 \in L^2(\mathbb{R}^N)$  and  $t > 0$ , we denote by  $S(t)u_0$  the function given by the right-hand side of (8.50). Verify that  $S(t)$  is a continuous linear operator from  $L^2(\mathbb{R}^N)$  into  $L^2(\mathbb{R}^N)$ . By setting  $S(0) = I$  (the identity of  $L^2(\mathbb{R}^N)$ ), verify that  $(S(t))_{t \geq 0}$  is a semigroup of operators which depend continuously on  $t$ , that is to say that they satisfy  $S(t + t') = S(t)S(t')$  for  $t, t' \geq 0$ . Let  $f \in C^1(\mathbb{R}^+; L^2(\mathbb{R}^N))$ . Show that the problem

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{in } ]0, +\infty[ \times \mathbb{R}^N \\ u(x, 0) = u_0(x) & \text{in } \mathbb{R}^N. \end{cases}$$

has a unique solution  $u \in C(\mathbb{R}^+; L^2(\mathbb{R}^N)) \cap C^1(\mathbb{R}_*^+; L^2(\mathbb{R}^N))$ , given by

$$u(t) = S(t)u_0 + \int_0^t S(t-s)f(s) ds,$$

that is to say

$$u(x, t) = \int_{\mathbb{R}^N} u_0(y) e^{-|x-y|^2/4t} \frac{dy}{(2\pi t)^{N/2}} + \int_0^t \int_{\mathbb{R}^N} f(y, s) e^{-|x-y|^2/4(t-s)} \frac{dy ds}{(2\pi(t-s))^{N/2}}.$$

The explicit formula (8.50) allows us to recover easily, for problem (8.49) posed in the whole space, the qualitative properties studied before. This is the object of the following exercises where we shall denote by  $u$  the solution (8.50) of the problem (8.49), with the initial data  $u_0 \in L^2(\mathbb{R}^N)$ .

**Exercise 8.4.7 (energy equality)** Show that, for all  $T > 0$ ,

$$\frac{1}{2} \int_{\mathbb{R}^N} u(x, T)^2 dx + \int_0^T \int_{\mathbb{R}^N} |\nabla u(x, t)|^2 dx dt = \frac{1}{2} \int_{\mathbb{R}^N} u_0(x)^2 dx.$$

**Exercise 8.4.8 (maximum principle)** Show that, if  $u_0 \in L^\infty(\mathbb{R}^N)$ , then  $u(t) \in L^\infty(\mathbb{R}^N)$  and

$$\|u(t)\|_{L^\infty(\mathbb{R}^N)} \leq \|u_0\|_{L^\infty(\mathbb{R}^N)} \quad \forall t > 0.$$

Show that, if  $u_0 \geq 0$  almost everywhere in  $\mathbb{R}^N$ , then  $u \geq 0$  in  $\mathbb{R}^N \times \mathbb{R}^+$ .

**Exercise 8.4.9 (regularizing effect)** Show that  $u \in C^\infty(\mathbb{R}^N \times \mathbb{R}_+^+)$ .

**Exercise 8.4.10 (asymptotic behaviour)** Show that

$$\lim_{|x| \rightarrow +\infty} u(x, t) = 0 \quad \forall t > 0 \quad \text{and} \quad \lim_{t \rightarrow +\infty} u(x, t) = 0 \quad \forall x \in \mathbb{R}^N.$$

**Exercise 8.4.11 (infinite speed of propagation)** Show that, if  $u_0 \geq 0$  and  $u_0 \not\equiv 0$ , then  $u(x, t) > 0$  in  $\mathbb{R}^N \times \mathbb{R}_+^+$ .

## 8.5 Qualitative properties in the hyperbolic case

### 8.5.1 Reversibility in time

We now examine the principal qualitative properties of the solution of the wave equation, which are very different from those of the solution of the heat flow equation. The most striking property, already stated in Chapter 1, is the **reversibility in time** of this equation.

**Proposition 8.5.1** *Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ , and take a final time  $T > 0$ . Take  $(v_0, v_1) \in H_0^1(\Omega) \times L^2(\Omega)$ , and a source term  $f \in L^2([0, T]; L^2(\Omega))$ . Then the **retrograde** wave equation (integrating backwards in time starting from  $T$ )*

$$\begin{cases} \frac{\partial^2 v}{\partial t^2} - \Delta v = f & \text{a.e. in } \Omega \times ]0, T[ \\ v = 0 & \text{a.e. on } \partial\Omega \times ]0, T[ \\ v(x, T) = v_0(x) & \text{a.e. in } \Omega \\ \frac{\partial v}{\partial t}(x, T) = v_1(x) & \text{a.e. in } \Omega \end{cases} \quad (8.52)$$

has a unique solution  $v \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ . Further, if  $u(t, x)$  is the solution of the wave equation (8.3) and if  $v_0(x) = u(x, T)$  in  $H_0^1(\Omega)$  and  $v_1(x) = \frac{\partial u}{\partial t}(x, T)$  in  $L^2(\Omega)$ , then we have  $v(t, x) = u(t, x)$ .

**Proof.** We make the change of unknown  $w(x, t) = v(x, T - t)$  and (8.52) becomes a ‘progressive’ wave equation with initial data at  $t = 0$  like the ‘usual’ equation (8.3) (as the time derivative is second order, there is no change in sign in the equation after this change of unknown). Applying theorem 8.3.4 (8.52) therefore has a unique solution. If  $v_0(x) = u(x, T)$  and  $v_1(x) = \frac{\partial u}{\partial t}(x, T)$ , the solution  $u(t, x)$  of (8.3) is also a solution of (8.52). By uniqueness we deduce  $v(t, x) = u(t, x)$ .  $\square$

The reversibility in time of the wave equation has numerous consequences. The most important is that there is no **regularizing effect** for the wave equation as

opposed to the case of the heat equation. Indeed, if this was the case, changing the sense of time like in proposition 8.5.1, we would obtain a contradictory ‘deregularizing’ effect (the solution will be less regular than that at the final time  $T$ , which is not possible since the regularizing effect must also apply). Consequently, **there is neither gain nor loss of regularity** for the solution of the wave equation with respect to the initial data. We can at most affirm that, as in the elliptic case, the regularity of the solution of the wave equation is directly linked to that of the data.

**Proposition 8.5.2** *Let  $\Omega$  be a regular open bounded set of  $\mathbb{R}^N$ . Take a final time  $T > 0$ , initial data  $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$  and  $u_1 \in H_0^1(\Omega)$ , a source term  $f \in L^2([0, T]; L^2(\Omega))$ , and  $u \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  the unique solution of the wave equation (8.33). Then,  $u$  belongs to  $C([0, T]; H_0^1(\Omega) \cap H^2(\Omega)) \cap C^1([0, T]; H_0^1(\Omega)) \cap C^2([0, T]; L^2(\Omega))$ .*

We will assume proposition 8.5.2 which is similar to proposition 8.4.6 that we have proved earlier. We can of course ‘increase’ regularity starting from this result and find that the solution  $u$  of the wave equation (8.33) is as regular as we want, provided that the data  $u_0$ ,  $u_1$  and  $f$  also are (with possible compatibility conditions on the data, see remark 8.4.7).

## 8.5.2 Asymptotic behaviour and equipartition of energy

**There is no maximum principle** for the wave equation. In the absence of a source term ( $f = 0$ ), even if the initial velocity is zero ( $u_1 = 0$ ) and if the initial data is positive ( $u_0 \geq 0$ ), the solution  $u$  can change sign in the course of time. This absence of maximum principle agrees with physical intuition. Let us imagine a cord or an elastic membrane: if we initially deform it in a position above its plane of rest, it will vibrate going alternately above and below this plane (in other words  $u$  changes sign). Mathematically, this counterexample can be written simply in the following form. Let  $w(x)$  be the first eigenfunction of the Laplacian in a connected bounded domain  $\Omega$  with Dirichlet boundary condition. From theorem 7.3.10, we can normalize  $w$  so that  $w(x) \geq 0$  in  $\Omega$ . Denoting by  $\lambda = \omega^2$  the first eigenvalue associated with  $w$ , it is easy to verify that  $u(t, x) = \cos(\omega t)w(x)$  changes sign in the course of time while being the unique solution in  $C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  of the wave equation (8.33) without source term and with the initial data

$$u(x, 0) = w(x), \quad \frac{\partial u}{\partial t}(x, 0) = 0 \quad \text{in } \Omega.$$

**There is therefore no asymptotic behaviour** in large time for the wave equation in a bounded domain. In other words, even if the source term  $f$  does not depend on time, the solution  $u$  does not converge to a stationary limit as the time  $t$  tends to infinity. In particular, if  $f = 0$ , the influence of the initial conditions is the same at all time since the energy is conserved and does not decrease (see exercise 8.3.1). The same counterexample  $u(t, x) = \cos(\omega t)w(x)$  allows us to see that there is no stationary limit of the oscillations which continue without damping.

This is obviously not the case for the damped wave equation (8.53) as the following exercise shows.

**Exercise 8.5.1** Let  $\eta > 0$ . We consider the damped wave equation

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} + \eta \frac{\partial u}{\partial t} - \Delta u = f & \text{in } \Omega \times \mathbb{R}_+^* \\ u = 0 & \text{on } \partial\Omega \times \mathbb{R}_+^* \\ u(x, 0) = u_0(x) & \text{in } \Omega \\ \frac{\partial u}{\partial t}(x, 0) = u_1(x) & \text{in } \Omega. \end{cases} \quad (8.53)$$

We assume that  $u$  is a sufficiently regular solution of (8.53) and that  $f$  is zero after a finite time. Show, with the help of Gronwall's lemma (see exercise 8.3.1), that  $u$  and  $\frac{\partial u}{\partial t}$  decrease exponentially to zero as time  $t$  tends to infinity.

Proposition 8.3.5 has established a property of conservation of total energy in the absence of a source term. A more precise result, called **equipartition of energy**, confirms that the total energy divides equally into kinetic energy and mechanical energy, asymptotically for large time.

**Exercise 8.5.2** Let  $u(t, x)$  be the solution, assumed to be sufficiently regular, of the wave equation (8.33). In the absence of a source term, show that

$$\lim_{t \rightarrow +\infty} \frac{1}{t} \int_0^t \int_{\Omega} \left| \frac{\partial u}{\partial t} \right|^2 dx = \lim_{t \rightarrow +\infty} \frac{1}{t} \int_0^t \int_{\Omega} |\nabla u|^2 dx = \frac{1}{2} E_0,$$

with  $E_0$  the initial energy.

$$E_0 = \int_{\Omega} |u_1(x)|^2 dx + \int_{\Omega} |\nabla u_0(x)|^2 dx.$$

For this we shall multiply the equation (8.33) by  $u$  and integrate by parts.

### 8.5.3 Finite velocity of propagation

One last important property of the wave equation is **propagation at finite velocity**. We have already seen in Chapter 1 (in  $N = 1$  dimension and in the entire space  $\Omega = \mathbb{R}$ ) that there exists a light cone (or domain of dependence) which contains all the information on the solution of the wave equation (see Figure 1.3). More precisely, the solution  $u$  in  $(t, x)$  only depends on the values of the initial data  $u_0$  and  $u_1$  on the segment  $[x - t, x + t]$ . We deduce that if the initial data have compact support  $K = [k_{\inf}, k_{\sup}] \subset \mathbb{R}$ , then the solution at time  $t$  has compact support in  $[k_{\inf} - t, k_{\sup} + t]$ . In physical terms this says that the initial perturbations propagate at finite velocity bounded by 1. This situation is again very different from what happens for the heat flow equation (compare with proposition 8.4.3). The following exercise allows us to generalize this discussion to  $N=2, 3$  dimensions.

**Exercise 8.5.3** We consider the wave equation in the entire space  $\mathbb{R}^N$

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta u = 0 & \text{in } \mathbb{R}^N \times \mathbb{R}_+^* \\ u(x, 0) = u_0(x) & \text{in } x \in \mathbb{R}^N \\ \frac{\partial u}{\partial t}(x, 0) = u_1(x) & \text{in } x \in \mathbb{R}^N, \end{cases} \quad (8.54)$$



with initial data  $(u_0, u_1)$  which is regular with compact support. Show that the solution  $u(t, x)$  can be put in the form

$$u(x, t) = (Mu_1)(x, t) + \left( \frac{\partial(Mu_0)}{\partial t} \right)(x, t),$$

where  $M$  is an averaging operator defined by

$$\begin{aligned} \text{if } N = 1, \quad (Mv)(x, t) &= \frac{1}{2} \int_{-t}^{+t} v(x - \xi) d\xi, \\ \text{if } N = 2, \quad (Mv)(x, t) &= \frac{1}{2\pi} \int_{|\xi| < t} \frac{v(x - \xi)}{\sqrt{t^2 - |\xi|^2}} d\xi, \\ \text{if } N = 3, \quad (Mv)(x, t) &= \frac{1}{4\pi t} \int_{|\xi|=t} v(x - \xi) ds(\xi), \end{aligned}$$

where  $ds(\xi)$  is the surface measure of the sphere. Deduce that the solution  $u$  in  $(t, x)$  depends only on the values of the initial data  $u_0$  and  $u_1$  on the ball  $|x| \leq t$ . (To know how to find the expressions above for the operator  $M$ , we refer to Chapter VII of [38].)

**Exercise 8.5.4** We consider the wave equation (8.54) in a domain  $\Omega \subset \mathbb{R}^N$  with Dirichlet or Neumann (homogeneous), boundary conditions and initial data  $(u_0, u_1)$  which is regular with compact support in  $\Omega$ . Verify that there exists a time  $T > 0$  such that on the interval  $[0, T]$  the solution is again given by the formulas of exercise 8.5.3.

The following exercise displays the essential difference between  $N = 2$  and 3 space dimensions.

**Exercise 8.5.5 (musical application)** Assuming that sound propagates according to the wave equation, show that it is not possible to listen to (audible) music in a world with  $N = 2$  space dimensions, but that this is (very happily) possible in  $N = 3$  dimensions.

## 8.6 Numerical methods in the parabolic case

In this section, we show how the finite element method (presented in Chapter 6) adapts to the numerical solution of the heat flow equation: we use finite elements for the spatial discretization, and finite differences for the temporal discretization.

### 8.6.1 Semidiscretization in space

We discretize the variational formulation (8.6) of the heat equation (8.1) **in space only**. For this, as in the case of elliptic problems, we construct an internal variational approximation by introducing a subspace  $V_{0h}$  of  $H_0^1(\Omega)$ , of finite dimension. Typically,

$V_{0h}$  will be a subspace of  $\mathbb{P}_k$  or  $\mathbb{Q}_k$  finite elements) on a triangular (or rectangular) mesh as made precise in the definitions 6.3.5 and 6.3.25.

The semidiscretization of (8.6) is therefore the following variational approximation: find  $u_h(t)$  a function of  $]0, T[$  with values in  $V_{0h}$  such that

$$\begin{cases} \frac{d}{dt} \langle u_h(t), v_h \rangle_{L^2(\Omega)} + a(u_h(t), v_h) = \langle f(t), v_h \rangle_{L^2(\Omega)} & \forall v_h \in V_{0h}, \quad 0 < t < T, \\ u_h(t=0) = u_{0,h} \end{cases} \quad (8.55)$$

where  $u_{0,h} \in V_{0h}$  is an approximation of the initial data  $u_0$ . This method of approximation is also known as the ‘method of lines’.

We can adapt the abstract framework of theorem 8.2.3 to show that (8.55) has a unique solution, but it is much more simple to verify that (8.55) is in fact a system of **ordinary differential equations** with constant coefficients from where we can easily calculate a unique solution. Practically, to solve (8.55) we introduce a basis  $(\phi_i)_{1 \leq i \leq n_{dl}}$  of  $V_{0h}$  (typically, a finite element basis), and we look for  $u_h(t)$  in the form

$$u_h(t) = \sum_{i=1}^{n_{dl}} U_i^h(t) \phi_i, \quad (8.56)$$

with  $U^h = (U_i^h)_{1 \leq i \leq n_{dl}}$  the vector of the coordinates of  $u_h$ . It is important to note that in (8.56) the basis functions  $\phi_i$  do not depend on time and that only the coordinates  $U_i^h(t)$  are functions of time  $t$ . Likewise, we set

$$u_{0,h} = \sum_{i=1}^{n_{dl}} U_i^{0,h} \phi_i,$$

and (8.55) becomes, for all  $1 \leq i \leq n_{dl}$ ,

$$\begin{cases} \sum_{j=1}^{n_{dl}} \langle \phi_j, \phi_i \rangle_{L^2(\Omega)} \frac{dU_j^h(t)}{dt} + \sum_{j=1}^{n_{dl}} a(\phi_j, \phi_i) U_j^h(t) = \langle f(t), \phi_i \rangle_{L^2(\Omega)} \\ U_i^h(t=0) = U_i^{0,h} \end{cases}$$

Introducing (as in the proof of lemma 7.4.1) the **mass matrix**  $\mathcal{M}_h$  defined by

$$(\mathcal{M}_h)_{ij} = \langle \phi_i, \phi_j \rangle_{L^2(\Omega)} \quad 1 \leq i, j \leq n_{dl},$$

and the **stiffness matrix**  $\mathcal{K}_h$  defined by

$$(\mathcal{K}_h)_{ij} = a(\phi_i, \phi_j) \quad 1 \leq i, j \leq n_{dl},$$

the variational approximation (8.55) is equivalent to the linear system of **ordinary differential equations** with constant coefficients

$$\begin{cases} \mathcal{M}_h \frac{dU^h}{dt}(t) + \mathcal{K}_h U^h(t) = b^h(t), & 0 < t < T, \\ U^h(t=0) = U^{0,h} \end{cases} \quad (8.57)$$

with  $b_i^h(t) = \langle f(t), \phi_i \rangle_{L^2(\Omega)}$ . The existence and the uniqueness, as well as an explicit formula, of the solution of (8.57) is obtained classically by simple simultaneous diagonalization of  $\mathcal{M}_h$  and  $\mathcal{K}_h$  (see Section 7.4.1 on this subject). As it is difficult and costly to diagonalize (8.57), in practice we solve (8.57) numerically by discretization and time stepping. There exist numerous classical methods for numerical calculation of the solutions of ordinary differential equations. We see some of these in the following section. Before that, we state a convergence result of the ‘semidiscrete’ solutions of (8.55) to the exact solution of (8.6).

**Proposition 8.6.1** *Let  $(\mathcal{T}_h)_{h>0}$  be a sequence of regular triangular meshes of  $\Omega$ . Let  $V_{0h}$  be the subspace of  $H_0^1(\Omega)$ , defined by the  $\mathbb{P}_k$  finite element method, of dimension  $n_{dl}$ . Take  $f(t) \in L^2([0, T]; L^2(\Omega))$ ,  $u_0 \in H_0^1(\Omega)$ , and  $u \in L^2([0, T]; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ , the unique solution of the heat flow equation (8.12). Let  $u_h$  be the unique solution of the variational approximation (8.55) in  $V_{0h}$ . If  $\lim_{h \rightarrow 0} \|u_{0,h} - u_0\|_{L^2(\Omega)} = 0$ , then we have*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{L^2([0, T]; H_0^1(\Omega))} = \lim_{h \rightarrow 0} \|u - u_h\|_{C([0, T]; L^2(\Omega))} = 0.$$

The proof of proposition 8.6.1 is similar to that of the preceding results on variational approximation. It can be found in the work [34]. We can also obtain error estimates and explicit rates of convergence, but this would involve too much.

## 8.6.2 Total discretization in space-time

After having discretized the heat flow equation in space by a finite element method, we finish the discretization of the problem by using **finite differences in time**. Concretely, we use finite differences schemes to solve the system of ordinary differential equations (8.57) resulting from the semidiscretization in space. We shall therefore revisit many schemes already studied in Chapter 2 as well as the notions such as stability or the order of precision. To simplify the notation, we rewrite the system (8.57) without mentioning the dependence on the parameter  $h$  of the spatial mesh

$$\begin{cases} \mathcal{M} \frac{dU}{dt}(t) + \mathcal{K}U(t) = b(t) \\ U(t=0) = U^0 \end{cases} \quad (8.58)$$

To simplify the analysis we shall assume that  $b(t)$  is continuous on  $[0, T]$ . We break the time interval  $[0, T]$  into  $n_0$  intervals or timesteps  $\Delta t = T/n_0$  and we set

$$t_n = n\Delta t \quad 0 \leq n \leq n_0.$$

We denote by  $U^n$  the approximation of  $U(t_n)$  calculated by a scheme. To calculate the approximate solutions of (8.58) numerically, the simplest and most frequently used scheme is the  $\theta$ -**scheme** (see (2.5))

$$\mathcal{M} \frac{U^{n+1} - U^n}{\Delta t} + \mathcal{K}(\theta U^{n+1} + (1-\theta)U^n) = \theta b(t_{n+1}) + (1-\theta)b(t_n). \quad (8.59)$$

When  $\theta = 0$ , we call (8.59) an **explicit scheme**, and when  $\theta = 1$ , an **implicit scheme**, and for  $\theta = 1/2$ , the **Crank–Nicolson scheme**. We can rewrite (8.59) in the form

$$(\mathcal{M} + \theta \Delta t \mathcal{K})U^{n+1} = (\mathcal{M} - (1 - \theta) \Delta t \mathcal{K})U^n + \Delta t(\theta b(t_{n+1}) + (1 - \theta)b(t_n)). \quad (8.60)$$

Let us remark that in general the matrix  $\mathcal{M}$  is not diagonal, and therefore, even for the explicit scheme, it is necessary to solve a linear system to calculate  $U^{n+1}$  as a function of  $U^n$ , and of the right-hand side  $b$  (except if we use a numerical integration formula which makes  $\mathcal{M}$  diagonal, see remark 7.4.3). Obviously, we can construct many schemes inspired from those in Chapter 2. Let us cite an example of a scheme with three time levels, the **Gear scheme**,

$$\mathcal{M} \frac{3U^{n+1} - 4U^n + U^{n-1}}{2\Delta t} + \mathcal{K}U^{n+1} = b(t_{n+1}). \quad (8.61)$$

These schemes are, of course, consistent (see definition 2.2.4) and we can easily analyse their precision (with respect to the time variable).

**Exercise 8.6.1** Show that the Crank–Nicolson scheme and the Gear scheme are second order (in time), while the  $\theta$ -scheme for  $\theta \neq 1/2$  is first order.

**Remark 8.6.2** It is possible to construct a finite element method **in space and in time**, but this is not of interest except in the case where the domain  $\Omega(t)$  varies as a function of time. •

We now give a definition of the stability of these schemes which is a variant of definition 2.2.8.

**Definition 8.6.3** A finite difference scheme for (8.58) is called *stable* if

$$\mathcal{M}U^n \cdot U^n \leq C \quad \text{for all } 0 \leq n \leq n_0 = T/\Delta t,$$

where the constant  $C > 0$  is independent of  $\Delta t$  and of the dimension of the system  $n_{dl}$  (therefore of the mesh size  $h$ ), but can depend on the initial data  $U^0$ , on the right-hand side  $b$ , and on  $T$ .

**Remark 8.6.4** The choice of the norm  $\sqrt{\mathcal{M}U \cdot U}$  in definition 8.6.3 is explained by the fact that  $\mathcal{M}U \cdot U = \int_{\Omega} |u|^2 dx$  with  $u \in V_{0h}$ , a function of the coordinates  $U$  in the chosen basis of  $V_{0h}$  ( $\mathcal{M}$  is positive definite). Let us recall that, in definition 2.2.8 of stability in the sense of the finite differences, we weighted the Euclidean norm of  $U$  by  $\Delta x$  to recover the link with the norm of  $u$  in  $L^2(\Omega)$ . •

**Lemma 8.6.5** If  $1/2 \leq \theta \leq 1$ , the  $\theta$ -scheme (8.59) is unconditionally stable, while, if  $0 \leq \theta < 1/2$ , it is stable under the CFL condition

$$\max_i \lambda_i \Delta t \leq \frac{2}{1 - 2\theta}, \quad (8.62)$$

where the  $\lambda_i$  are the eigenvalues of  $\mathcal{K}U = \lambda \mathcal{M}U$  (see (7.23)).

**Remark 8.6.6** We do not immediately recognize in (8.62) the usual CFL (Courant–Friedrichs–Lewy) condition  $\Delta t \leq Ch^2$  for the heat equation (see Section 2.2.3). In fact, we can show that, if the mesh  $\mathcal{T}_h$  is uniformly regular in the sense where every element contains a ball of radius  $Ch$  (with  $C > 0$  independent of the element), then we effectively have  $\max_i \lambda_i = \mathcal{O}(h^{-2})$ . We suggest that the reader verifies this fact in  $N = 1$  dimension as exercise 8.6.4 below. In practice we do not use the  $\theta$ -scheme for  $\theta < 1/2$  since the stability condition (8.62) is much too severe: it needs the use of very small timesteps which makes the calculation much too costly. •

**Proof.** We rewrite the scheme (8.60) in the orthonormal basis for  $\mathcal{M}$  and orthogonal for  $\mathcal{K}$  (see the proof of lemma 7.4.1)

$$(\mathbf{I} + \theta \Delta t \operatorname{diag}(\lambda_i)) \tilde{U}^{n+1} = (\mathbf{I} - (1 - \theta) \Delta t \operatorname{diag}(\lambda_i)) \tilde{U}^n + \Delta t \tilde{b}^n, \quad (8.63)$$

with  $\mathcal{M} = PP^*$ ,  $\mathcal{K} = P \operatorname{diag}(\lambda_i) P^*$ ,  $\tilde{U}^n = P^* U^n$ , and  $\tilde{b}^n = P^{-1}(\theta b(t_{n+1}) + (1 - \theta)b(t_n))$ . We deduce from (8.63) that the components  $\tilde{U}_i^n$  of  $\tilde{U}^n$  satisfy the following equation:

$$\tilde{U}_i^n = (\rho_i)^n \tilde{U}_i^0 + \frac{\Delta t}{1 + \theta \Delta t \lambda_i} \sum_{k=1}^n (\rho_i)^{k-1} \tilde{b}_i^{n-k}. \quad (8.64)$$

with

$$\rho_i = \frac{1 - (1 - \theta) \Delta t \lambda_i}{1 + \theta \Delta t \lambda_i}.$$

In this basis the stability condition is  $\|U^n\|_{\mathcal{M}} = \|\tilde{U}^n\| \leq C$ . Consequently, a necessary and sufficient condition for stability is  $|\rho_i| \leq 1$  for all  $i$ , which is none other than the condition (8.62) if  $0 \leq \theta < 1/2$ , and which is always satisfied if  $\theta \geq 1/2$ . □

**Remark 8.6.7** It is clear in the estimate (8.64) that the larger the value of  $\theta$ , the smaller the coefficient in front of the term  $\tilde{b}_i^{n-k}$ . In fact, this property corresponds to an exponential damping by the scheme of the contributions from the source term. Consequently, even if for every value  $1/2 < \theta \leq 1$  the  $\theta$ -scheme is stable, its maximum stability is attained for  $\theta = 1$  (the numerical errors decay). This is why the implicit scheme is more robust and often used for ‘stiff’ problems even though it is less precise than the Crank–Nicolson scheme. •

**Remark 8.6.8** The system of ordinary differential equations (8.58) is called ‘stiff’ since its solutions involve terms of the type  $\exp(-\lambda_i t)$ , which evolve on very different time scales (let us recall that  $\min_i \lambda_i = \mathcal{O}(1)$  and  $\max_i \lambda_i = \mathcal{O}(h^{-2})$ ). There exist other numerical integration methods for ordinary differential equations which are more powerful and more complicated than the  $\theta$ -scheme. We cite the Runge–Kutta methods (see, for example, the Chapter XX of [14]). Nevertheless, there is a compromise to be struck between the use of a numerical integration method in time which is robust, but expensive, and the (very important) size of the systems to be solved (let us recall that the size is proportional to the number of elements). •

We can use the variational character of the discretization by finite elements which has led to (8.58) to show the unconditional stability of the  $\theta$ -scheme in another way.

**Exercise 8.6.2** We consider the  $\theta$ -scheme (8.59) with  $1/2 \leq \theta \leq 1$ . We denote by  $\|U\|_{\mathcal{M}} = \sqrt{\mathcal{M}U \cdot U}$ . Prove the following discrete equivalent of the energy inequality (8.17)

$$\|U^{n_0}\|_{\mathcal{M}}^2 + \Delta t \sum_{n=0}^{n_0} \mathcal{K} \hat{U}^n \cdot \hat{U}^n \leq C \left( \|U^0\|_{\mathcal{M}}^2 + \int_0^T \|f(t)\|_{L^2(\Omega)}^2 dt + \mathcal{O}(1) \right).$$

For this, we will take the scalar product of (8.59) with  $\hat{U}^n = \theta U^{n+1} + (1 - \theta)U^n$ .

**Exercise 8.6.3** Show that the Gear scheme (8.61) is unconditionally stable.

**Exercise 8.6.4** We solve the heat flow equation (8.12) in  $N = 1$  dimension by  $\mathbb{P}_1$  finite elements and the explicit scheme in time. We use a quadrature formula which makes the matrix  $\mathcal{M}$  diagonal (see remark 7.4.3 and exercise 7.4.1). We recall that the matrix  $\mathcal{K}$  is given by (6.12) and that we have calculated its eigenvalues in exercise 13.1.3. Show that in this case the CFL condition (8.62) is of the type  $\Delta t \leq Ch^2$ .

Finally, we can state a convergence result for this method of discretization that we will not prove (the proof of proposition 8.6.9 is in the spirit of the preceding proofs). For more details, as well as error estimates and explicit rates of convergence, we refer to [34].

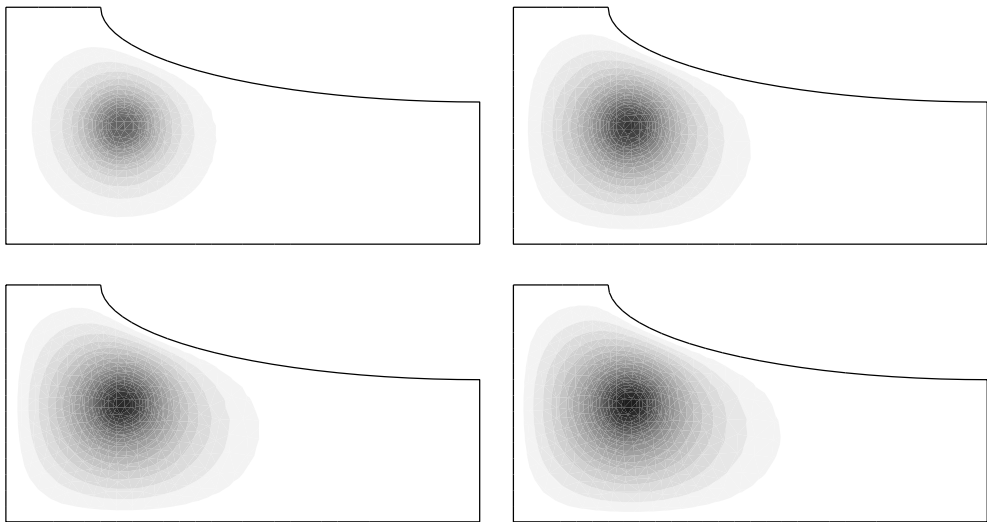
**Proposition 8.6.9** *Let  $u$  be the ‘sufficiently regular’ solution of the heat equation (8.12). Let  $(\mathcal{T}_h)_{h>0}$  be a sequence of regular triangular meshes of  $\Omega$ . Let  $V_{0h}$  be the subspace of  $H_0^1(\Omega)$ , defined by the  $\mathbb{P}_k$  finite element method. Let  $(\Delta t)$  be a sequence of timesteps which tends to zero. Let  $u_h^n \in V_{0h}$  be the function whose coordinates  $U^n$  in the finite element basis of  $V_{0h}$  are calculated by the  $\theta$ -scheme. If  $\lim_{h \rightarrow 0} u_h^0 = u_0$  in  $L^2(\Omega)$ , and if  $h$  and  $\Delta t$  tend to 0, respecting the stability condition (8.62), then we have*

$$\lim_{h \rightarrow 0} \max_{\Delta t \rightarrow 0, 0 \leq n \leq n_0} \|u(t_n) - u_h^n\|_{L^2(\Omega)} = 0.$$

To finish we illustrate this by taking the example of Section 6.3 on the diffusion in the atmosphere of a pollutant emitted by a localized source (see the mesh in Figure 6.7 and the right-hand side  $f$  in Figure 6.14). The initial data  $u_0$  is taken zero in the domain. Figure 8.1 presents the results at 4 different times. As the source term is independent of time, the solution converges towards a stationary regime as the time tends to infinity.

## 8.7 Numerical methods in the hyperbolic case

The numerical methods to solve the wave equation are very similar (in principle, but not always in practice) to those that we have seen for the heat equation.

Figure 8.1. Concentration at time  $t = 1, 2, 3, 5$ .

### 8.7.1 Semidiscretization in space

We discretize **in space only** the variational formulation (8.25) of the wave equation (8.3). For this, we construct an internal variational approximation by introducing a subspace  $V_{0h}$  of  $H_0^1(\Omega)$ , of finite dimension (typically, a finite element subspace). The semidiscretization of (8.25) is therefore the following variational approximation: find  $u_h(t)$  a function of  $]0, T[$  with values in  $V_{0h}$  such that

$$\begin{cases} \frac{d^2}{dt^2} \langle u_h(t), v_h \rangle_{L^2(\Omega)} + a(u_h(t), v_h) = \langle f(t), v_h \rangle_{L^2(\Omega)} & \forall v_h \in V_{0h}, \quad 0 < t < T, \\ u_h(t=0) = u_{0,h}, \quad \frac{\partial u_h}{\partial t}(t=0) = u_{1,h} \end{cases} \quad (8.65)$$

where  $u_{0,h} \in V_{0h}$  and  $u_{1,h} \in V_{0h}$  are approximations of the initial data  $u_0$  and  $u_1$ .

To show that (8.65) has a unique solution and the practical way to calculate it, we introduce a basis  $(\phi_i)_{1 \leq i \leq n_{dl}}$  of  $V_{0h}$  (which does not depend on time), and we look for  $u_h(t)$  in the form

$$u_h(t) = \sum_{i=1}^{n_{dl}} U_i^h(t) \phi_i,$$

with  $U^h = (U_i^h)_{1 \leq i \leq n_{dl}}$  the vector of the coordinates of  $u_h$ . By setting

$$u_{0,h} = \sum_{i=1}^{n_{dl}} U_i^{0,h} \phi_i, \quad u_{1,h} = \sum_{i=1}^{n_{dl}} U_i^{1,h} \phi_i, \quad b_i^h(t) = \langle f(t), \phi_i \rangle_{L^2(\Omega)}, \quad 1 \leq i \leq n_{dl},$$

the variational approximation (8.65) is equivalent to the linear system of second order **ordinary differential equations** with constant coefficients

$$\begin{cases} \mathcal{M}_h \frac{d^2 U^h}{dt^2}(t) + \mathcal{K}_h U^h(t) = b^h(t), & 0 < t < T, \\ U^h(t=0) = U^{0,h}, & \frac{dU^h}{dt}(t=0) = U^{1,h}, \end{cases} \quad (8.66)$$

where we again find the **mass matrix**  $\mathcal{M}_h$  and **stiffness matrix**  $\mathcal{K}_h$  as for the heat flow equation

$$(\mathcal{M}_h)_{ij} = \langle \phi_i, \phi_j \rangle_{L^2(\Omega)}, \quad (\mathcal{K}_h)_{ij} = a(\phi_i, \phi_j) \quad 1 \leq i, j \leq n_{dl}.$$

The existence and uniqueness, as well as an explicit formula, of the solution of (8.66) is easily obtained by simple simultaneous diagonalization of the matrices  $\mathcal{M}_h$  and  $\mathcal{K}_h$ . As it is difficult and costly to diagonalize (8.66), in practice we solve (8.66) numerically by discretization and time stepping.

**Exercise 8.7.1** Write the linear system of ordinary differential equations obtained by semidiscretization of the damped wave equation (8.53).

## 8.7.2 Total discretization in space-time

We use a method of **finite differences in time** to solve the system of ordinary differential equations (8.66). To simplify the notation, we rewrite the system (8.66) without mentioning the spatial dependence on  $h$

$$\begin{cases} \mathcal{M} \frac{d^2 U}{dt^2}(t) + \mathcal{K} U(t) = b(t) \\ U(t=0) = U_0, & \frac{dU}{dt}(t=0) = U_1, \end{cases} \quad (8.67)$$

where we assume that  $b(t)$  is continuous on  $[0, T]$ . We break the time interval  $[0, T]$  into  $n_0$  timesteps  $\Delta t = T/n_0$ , we set  $t_n = n\Delta t$   $0 \leq n \leq n_0$ , and we denote by  $U^n$  the approximation of  $U(t_n)$  calculated by a scheme. For  $0 \leq \theta \leq 1/2$  we propose the  **$\theta$ -scheme**

$$\begin{aligned} & \mathcal{M} \frac{U^{n+1} - 2U^n + U^{n-1}}{(\Delta t)^2} + \mathcal{K} (\theta U^{n+1} + (1 - 2\theta)U^n + \theta U^{n-1}) \\ &= \theta b(t_{n+1}) + (1 - 2\theta)b(t_n) + \theta b(t_{n-1}). \end{aligned} \quad (8.68)$$

When  $\theta = 0$ , we call (8.68) an **explicit scheme** (it is only truly explicit if the mass matrix  $\mathcal{M}$  is diagonal). To start the scheme we must know  $U^0$  and  $U^1$ , which we obtain, thanks to the initial conditions

$$U^0 = U_0 \quad \text{and} \quad \frac{U^1 - U^0}{\Delta t} = U_1.$$



A more frequently used scheme, as it is more general, is the **Newmark scheme**. To solve the ‘damped’ system

$$\mathcal{M} \frac{d^2 U}{dt^2}(t) + \mathcal{C} \frac{dU}{dt}(t) + \mathcal{K}U(t) = b(t)$$

we approximate  $U(t)$ ,  $dU/dt(t)$ ,  $d^2U/dt^2(t)$  by three sequences  $U^n, \dot{U}^n, \ddot{U}^n$

$$\begin{cases} \mathcal{M}\ddot{U}^{n+1} + \mathcal{C}\dot{U}^{n+1} + \mathcal{K}U^{n+1} = b(t_{n+1}) \\ \dot{U}^{n+1} = \dot{U}^n + \Delta t(\delta\ddot{U}^{n+1} + (1-\delta)\ddot{U}^n) \\ U^{n+1} = U^n + \Delta t\dot{U}^n + \frac{(\Delta t)^2}{2} \left( 2\theta\ddot{U}^{n+1} + (1-2\theta)\ddot{U}^n \right) \end{cases} \quad (8.69)$$

with  $0 \leq \delta \leq 1$  and  $0 \leq \theta \leq 1/2$ . When the damping matrix is zero ( $\mathcal{C} = 0$ ), we can eliminate the sequences  $\dot{U}^n$  and  $\ddot{U}^n$ , and (8.69) is equivalent to

$$\begin{aligned} \mathcal{M} \frac{U^{n+1} - 2U^n + U^{n-1}}{(\Delta t)^2} + \mathcal{K} \left( \theta U^{n+1} + \left( \frac{1}{2} + \delta - 2\theta \right) U^n + \left( \frac{1}{2} - \delta + \theta \right) U^{n-1} \right) \\ = \theta b(t_{n+1}) + \left( \frac{1}{2} + \delta - 2\theta \right) b(t_n) + \left( \frac{1}{2} - \delta + \theta \right) b(t_{n-1}). \end{aligned} \quad (8.70)$$

Let us remark that for  $\delta = 1/2$  the Newmark scheme becomes the  $\theta$ -scheme. In practice, the larger of  $\delta$ , the more dissipative and robust the scheme (the numerical errors decay more quickly), even if it is less accurate.

**Exercise 8.7.2** Show that the Newmark scheme is first order (in time) for  $\delta \neq 1/2$ , second order for  $\delta = 1/2$  and  $\theta \neq 1/12$ , and fourth order if  $\delta = 1/2$  and  $\theta = 1/12$  (we limit ourselves to the equation without damping).

We study the stability of these schemes in the sense of definition 8.6.3. To avoid difficult calculation, we shall be content with studying the Von Neumann necessary stability condition (see remark 2.2.24). The following result is in the same spirit as lemma 2.3.6.

**Lemma 8.7.1** *We consider the Newmark scheme (8.70). If  $\delta < 1/2$ , it is always unstable. Let us assume from now on that  $\delta \geq 1/2$ . The Von Neumann necessary stability condition is always satisfied if  $\delta \leq 2\theta \leq 1$ , while, if  $0 \leq 2\theta < \delta$  it is only satisfied under the CFL condition*

$$\max_i \lambda_i (\Delta t)^2 < \frac{2}{\delta - 2\theta}, \quad (8.71)$$

where the  $\lambda_i$  are the eigenvalues of  $\mathcal{K}U = \lambda\mathcal{M}U$  (see (7.23)).

**Remark 8.7.2** We do not recognize the usual CFL (Courant–Friedrichs–Lewy) condition immediately in (8.71) for the wave equation (see lemma 2.3.6). In fact, we can show that, if the mesh  $\mathcal{T}_h$  is uniformly regular in the sense where every element contains a ball of radius  $Ch$  (with  $C > 0$  independent of the element), then we have effectively  $\max_i \lambda_i = \mathcal{O}(h^{-2})$ . The reader has checked this fact in  $N = 1$  dimension in exercise 8.6.4. Conversely to the parabolic case, the CFL condition (8.71) is not too severe since we can take the timesteps  $\Delta t$  of the order of the space step  $h$ . However, since we must invert a linear system (to be able to calculate  $U^{n+1}$  as a function of  $U^n, U^{n-1}$ , and of the right-hand side), there is no additional cost in using the Newmark scheme for values of  $\delta$  and  $\theta$  such that it is unconditionally stable. The only interesting case for a stable scheme under the CFL condition is the case where it is **explicit**, that is to say that there is no linear system to solve at each timestep. In effect, an explicit scheme needs very few operations per timestep and therefore leads to less costly calculations. The only possibility for the Newmark scheme (8.70) to be explicit is that  $\theta = 0$  and that the mass matrix  $\mathcal{M}$  is diagonal thanks to a numerical integration formula (see remark 7.4.3). This explicit scheme is often used in practice with  $\delta = 1/2$ . •

**Proof.** It is very similar to that of lemma 8.6.5 but with the technical complications which recall lemma 2.3.6. We decompose  $U^n$  and the right-hand side of (8.70) in the basis orthonormal for  $\mathcal{M}$  and orthogonal for  $\mathcal{K}$ . Consequently, (8.70) is equivalent, component by component, to

$$\frac{U_i^{n+1} - 2U_i^n + U_i^{n-1}}{(\Delta t)^2} + \lambda_i \left( \theta U_i^{n+1} + \left( \frac{1}{2} + \delta - 2\theta \right) U_i^n + \left( \frac{1}{2} - \delta + \theta \right) U_i^{n-1} \right) = b_i^n, \quad (8.72)$$

with the obvious notation (the  $\lambda_i$  are the eigenvalues of the matrix system  $\mathcal{K}V_i = \lambda_i \mathcal{M}V_i$ ). As the scheme (8.72) is three level (see Section 2.2.5), we introduce an iteration matrix  $A_i$  such that

$$\begin{pmatrix} U_i^{n+1} \\ U_i^n \end{pmatrix} = A_i \begin{pmatrix} U_i^n \\ U_i^{n-1} \end{pmatrix} + \frac{(\Delta t)^2}{1 + \theta \lambda_i (\Delta t)^2} \begin{pmatrix} b_i^n \\ 0 \end{pmatrix} \quad \text{with } A_i = \begin{pmatrix} a_{11} & a_{12} \\ 1 & 0 \end{pmatrix},$$

$$a_{11} = \frac{2 - \lambda_i (\Delta t)^2 ((1/2) + \delta - 2\theta)}{1 + \theta \lambda_i (\Delta t)^2}, \quad a_{12} = -\frac{1 + \lambda_i (\Delta t)^2 ((1/2) - \delta + \theta)}{1 + \theta \lambda_i (\Delta t)^2}.$$

We deduce that

$$\begin{pmatrix} U_i^{n+1} \\ U_i^n \end{pmatrix} = A_i^n \begin{pmatrix} U_i^1 \\ U_i^0 \end{pmatrix} U_i^0 + \frac{(\Delta t)^2}{1 + \theta \lambda_i (\Delta t)^2} \sum_{p=0}^{n-1} A_i^p \begin{pmatrix} b_i^{n-p} \\ 0 \end{pmatrix}. \quad (8.73)$$

The Von Neumann necessary stability condition is thus  $\rho(A_i) \leq 1$ . We therefore calculate the eigenvalues of  $A_i$  which are the roots of the following polynomial in  $\mu$

$$\mu^2 - a_{11}\mu - a_{12} = 0$$

whose discriminant is

$$\Delta = \frac{-4\lambda_i(\Delta t)^2 + \lambda_i^2(\Delta t)^4(((1/2) + \delta)^2 - 4\theta)}{(1 + \theta\lambda_i(\Delta t)^2)^2}.$$

We verify easily that the roots of this polynomial have modulus less than or equal to 1 if and only if we are in one of the two following cases: either  $\Delta \leq 0$  and  $a_{12} \geq -1$ , or  $\Delta > 0$  and  $1 - a_{12} \geq |a_{11}|$ . A tedious but simple calculation leads to the condition (8.71) (see if necessary the theorem 6, section 3, chapter XX in [14]).  $\square$

**Exercise 8.7.3** We consider the limit case of lemma 8.7.1, that is to say  $\delta = 1/2$  and  $\lambda_i(\Delta t)^2 = 4/(1 - 4\theta)$ . Show that the Newmark scheme is unstable in this case by showing that

$$A_i = \begin{pmatrix} -2 & -1 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad A_i^n = (-1)^n \begin{pmatrix} n+1 & n \\ -n & 1-n \end{pmatrix}.$$

remark that this is a 'weak' instability since the growth of  $A_i^n$  is linear and not exponential.

**Remark 8.7.3** In the absence of a source term the solutions of the system of ordinary differential equations (8.67) are oscillating functions of the type  $\cos(\omega_i t)$ , with  $\omega_i^2 = \lambda_i$ . As the numerical integration schemes of (8.67) have the tendency to damp these oscillations a little (we say that they are diffusive or dissipative; see the discussion after definition 2.3.5), we can prefer another method of solution of (8.67) based on a decomposition of the solution into the eigenvectors  $V_i$  of the matrix system  $\mathcal{K}V_i = \lambda_i \mathcal{M}V_i$ . We look for the solution  $U(t)$  of (8.67) in the form

$$U(t) = \sum_{i \in I} x_i(t) V_i,$$

where  $I$  is a collection of indices of eigenmodes chosen to represent the solution correctly, and  $x_i(t)$  is the solution of a scalar ordinary differential equation

$$\frac{d^2 x_i}{dt^2}(t) + \lambda_i x_i(t) = b_i(t)$$

that we can integrate easily, quickly, and precisely by the numerical method of choice. This method, called **superposition of modes**, is often used in vibration mechanics. It has the advantage of being not very dissipative, that is to say the oscillations are damped very little. Of course, the quality of the results depends in a large part on the choice of the modes in the decomposition, a choice often motivated by physical considerations.  $\bullet$

Finally, we can state a convergence result for this method of discretization that we shall not prove (the proof of proposition 8.7.4 is in the spirit of the preceding proofs). For more detail, as well as error estimates and explicit rates of convergence, we refer to [34].

**Proposition 8.7.4** *Let  $u$  be the ‘sufficiently regular’ solution of the wave equation (8.33). Let  $(\mathcal{T}_h)_{h>0}$  be a sequence of regular triangular meshes of  $\Omega$ . Let  $V_{0h}$  be the subspace of  $H_0^1(\Omega)$ , defined by the  $\mathbb{P}_k$  finite element method. Let  $(\Delta t)$  be a sequence of timesteps which tend to zero. Let  $u_h^n \in V_{0h}$  be the function whose coordinates  $U^n$  in the finite element basis of  $V_{0h}$  are calculated by the Newmark scheme. If  $\lim_{h \rightarrow 0} u_h^0 = u_0$  in  $L^2(\Omega)$ ,  $\lim_{h \rightarrow 0} u_h^1 = u_1$  in  $L^2(\Omega)$ , and if  $h$  and  $\Delta t$  tend to 0 respecting the stability condition (8.71), then we have*

$$\lim_{h \rightarrow 0, \Delta t \rightarrow 0} \max_{0 \leq n \leq n_0} \|u(t_n) - u_h^n\|_{L^2(\Omega)} = 0.$$

To finish, we illustrate this by simulating the propagation of a spherical wave in a square cavity with a reflecting boundary. We therefore solve the wave equation in  $\Omega = ]0, 1[^2$  with a Neumann boundary condition, a zero initial velocity and an initial displacement with compact support and spherical symmetry centred on the point  $(0.3, 0.4)$ . In Figure 8.2 we trace the modulus of the deformation  $\nabla u$  at the initial instant and at five later instants (this type of image, called a Schlieren diagram, represents what we would see in a real experiment).

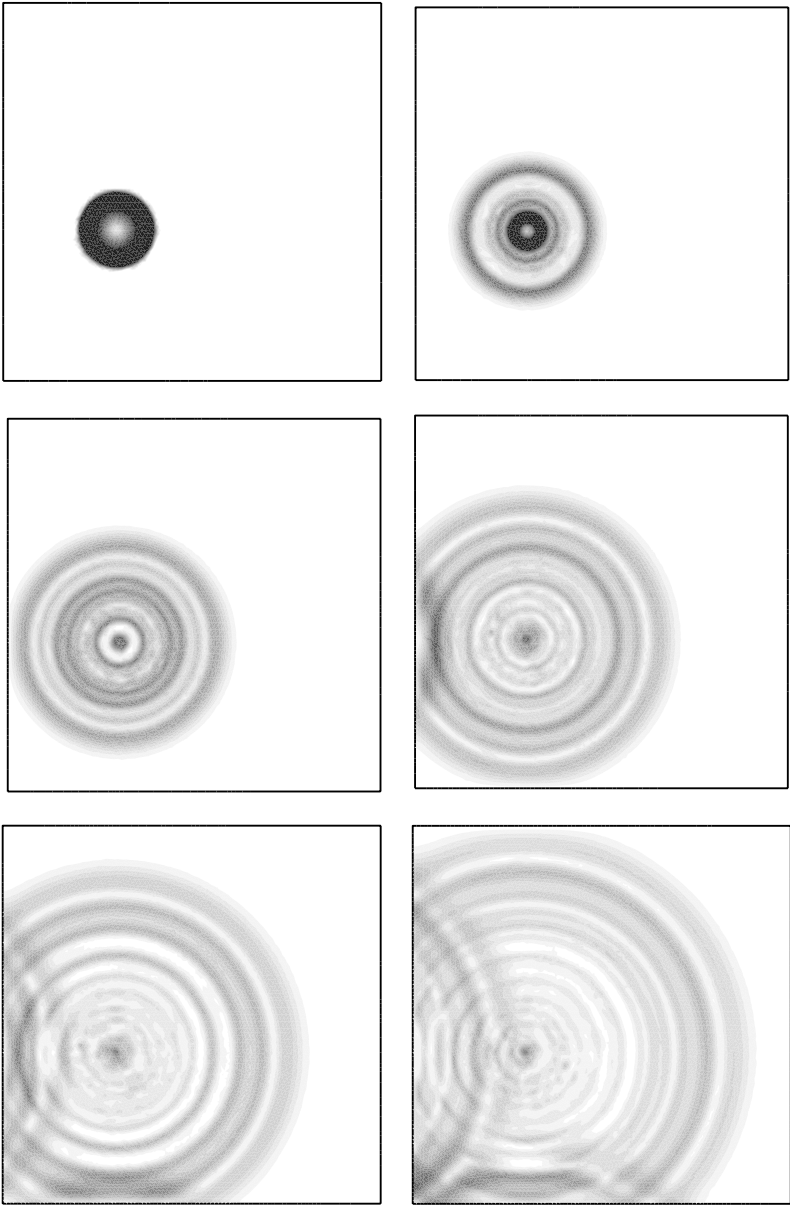


Figure 8.2. Modulus at time  $t = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5$  of  $|\nabla u|$ .

# 9 Introduction to optimization

---

## 9.1 Motivation and examples

### 9.1.1 Introduction

Optimization is a very old subject which has shown a resurgence since the appearance of computers and whose methods are applied in numerous domains: economics, management, planning, logistics, robotics, optimal design, engineering, signal processing, etc. Optimization is also a vast subject which touches on calculus of variations, operations research (optimization of management or decision processes), and optimal control. We will only mention these subjects since we would need complete notes for each of them if we want to treat them thoroughly.

In a certain way, optimization can be seen as a discipline independent of the numerical analysis of partial differential equations that we have studied in the preceding chapters. However, the interactions between these two disciplines are numerous and fertile and it is much more natural to deal with them in the same course. Indeed, after the **modelling** step from a physical phenomenon or an industrial system (possibly with the help of partial differential equations), and after the **numerical simulation** step, the work of the applied mathematician (who may be an engineer or researcher) is not finished: it is often necessary to **change** the system to improve certain aspects of it. This third step is that of **optimization**, that is to say that of the minimization (or the maximization) of a function which depends on the solution of the model.

In what follows we shall therefore mix examples of optimization problems where the models have a very different nature. In the simplest case, the model will be a simple algebraic equation and it will be a question of optimizing a function defined on a finite dimensional space (say  $\mathbb{R}^n$ ). Typically, this is the most frequent situation in operations research. A second category of problems corresponds to the case where the function

to optimize depends on the solution of an ordinary differential equation (in other words, this function is defined on an infinite dimensional space, say  $C[0, T]$ ). We then talk of optimal command, and the applications are numerous in robotics. The third and last category corresponds to the optimization of functions of the solution of a partial differential equation. This then involves the theory of optimal control of distributed systems which has numerous applications, for example, in optimal design or for the stabilization of mechanical structures. The next section displays several typical examples of these optimization problems. Let us remark that these categories are not hermetically sealed since after spatial and/or temporal discretization an ordinary or partial differential equation leads to a system of algebraic equations.

We can also separate optimization into two large branches of very different methods depending on whether the variables are continuous or discrete. Typically, if we minimize a function  $f(x)$  with  $x \in \mathbb{R}^n$ , we have **continuous optimization**, while if  $x \in \mathbb{Z}^n$  we have **combinatorial optimization** or discrete optimization. In spite of appearances, continuous optimization is often ‘easier’ than discrete optimization since we can use the idea of a derivative which is very useful from the theoretical as well as algorithmic point of view. Combinatorial optimization is natural and essential in many problems in operations research. This is a domain where, beside rigorous theoretical results, there are numerous ‘heuristic’ methods essential to obtaining good algorithmic performance.

To finish this brief introduction we describe the plan of the remainder. This chapter will principally consider the question of existence and uniqueness in continuous optimization, whether it is in finite or infinite dimensions. In particular, we see the crucial role of the **convexity** in obtaining existence results in infinite dimensions. Chapter 10 will develop the optimality conditions and the numerical algorithms which follow. Chapter 11 comprises an introduction to the methods of operations research, including linear programming and combinatorial methods. For more detail on optimization we refer the reader to the works [3], [8], [11], [30], [16], [19], [31].

### 9.1.2 Examples

We shall review several typical optimization problems, of unequal practical or theoretical importance, but which allow us to explore the different ‘branches’ of optimization.

Let us start with several examples in **operations research**, that is to say in optimization of the management or allocation of resources.

**Example 9.1.1 (transport problem)** This is an example of linear programming. The aim is to optimize the delivery of goods (a classical problem in logistics). We have  $M$  warehouses, indexed by  $1 \leq i \leq M$ , each one with a stock level  $s_i$ . We must deliver to  $N$  customers, indexed by  $1 \leq j \leq N$ , who have each ordered a quantity  $r_j$ . The unit cost of transport between warehouse  $i$  and customer  $j$  is given by  $c_{ij}$ . The decision variables are the quantities  $v_{ij}$  of goods leaving the warehouse  $i$  to the customer  $j$ . We want to minimize the cost of transport while satisfying the orders of

the customers (we assume that  $\sum_{i=1}^M s_i \geq \sum_{j=1}^N r_j$ ). In other words, we want to solve

$$\inf_{(v_{ij})} \left( \sum_{i=1}^M \sum_{j=1}^N c_{ij} v_{ij} \right)$$

under the constraints of stock limits and customer satisfaction

$$v_{ij} \geq 0, \quad \sum_{j=1}^N v_{ij} \leq s_i, \quad \sum_{i=1}^M v_{ij} = r_j \quad \text{for } 1 \leq i \leq M, \quad 1 \leq j \leq N.$$

This problem is a particular case of a transport problem. •

**Example 9.1.2 (assignment problem)** This is an example of combinatorial optimization or optimization in integer variables. Imagine yourself to be the head of a marriage agency. Take  $N$  women, indexed by  $1 \leq i \leq N$ , and  $N$  men, indexed by  $1 \leq j \leq N$ . If the woman  $i$  and the man  $j$  are suitable to be married their suitability variable  $a_{ij}$  is 1; if not it is 0. Let us remain classical: only heterosexual marriages are allowed and polygamy is not admitted. (We shall see in Section 11.3.7 that this hypothesis of monogamy is not necessary!) The aim is to maximize the number of marriages between these  $N$  women and  $N$  men. In other words, we look for a permutation  $\sigma$  in the set of permutations  $\mathcal{S}_N$  of  $\{1, \dots, N\}$  which realizes the maximum of

$$\max_{\sigma \in \mathcal{S}_N} \sum_{i=1}^N a_{i\sigma(i)}.$$

A variant consists of assigning values of  $a_{ij}$  between 0 and 1. This type of problem is called an assignment problem (it occurs in more serious industrial contexts such as the assignment of crews and aeroplanes in an airline company). Although it is not necessarily the best way to pose the problem, we can write it in a form close to example 9.1.1. The decision variables are denoted as  $v_{ij}$  which are 1 if there is a marriage between the woman  $i$  and the man  $j$  and 0 otherwise. We want to maximize

$$\sup_{(v_{ij})} \left( \sum_{i=1}^N \sum_{j=1}^N a_{ij} v_{ij} \right)$$

under the constraints

$$v_{ij} = 0 \text{ or } 1, \quad \sum_{j=1}^N v_{ij} \leq 1, \quad \sum_{i=1}^M v_{ij} \leq 1 \quad \text{for } 1 \leq i, \quad j \leq N.$$

We might believe that the assignment problem is simple since as there are a finite number of possibilities it is 'sufficient' to enumerate them to find the optimum. This is of course a delusion, since the characteristic of combinatorial problems is their



very large number of possible combinations, which in practice prevents an exhaustive enumeration. Nevertheless, for this problem there exist efficient solution techniques, see Section 11.3.7). •

**Example 9.1.3 (knapsack problem)** A classical problem is the knapsack problem. Take  $n$  objects with respective weight  $p_1, \dots, p_n \in \mathbb{R}$ , and respective utility  $u_1, \dots, u_n \in \mathbb{R}$ , and  $P \in \mathbb{R}$  a maximal weight that we want to carry. We set  $x_i = 1$  if we put the  $i$ th object in the knapsack, and  $x_i = 0$  otherwise. We want to maximize the utility of the knapsack under the weight constraint:

$$\max_{x \in \{0,1\}^n} \sum_{1 \leq i \leq n} x_i u_i.$$

$$\sum_{1 \leq i \leq n} x_i p_i \leq P$$

Again the difficulty for this is that the optimization variables  $x_i$  are discrete (see exercise 11.4.5 for a solution method). •

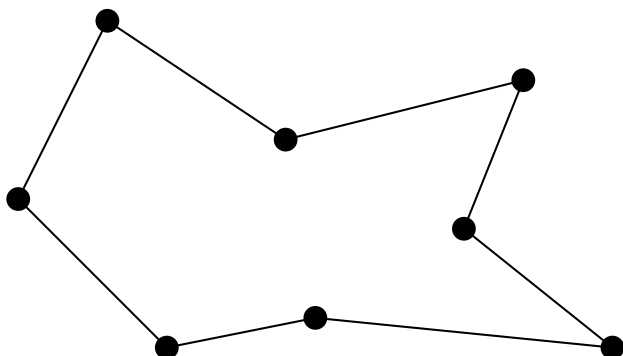


Figure 9.1. Route for a travelling salesman in example 9.1.4.

**Example 9.1.4 (travelling salesman problem)** A celebrated example in combinatorial optimization is the problem of the travelling salesman. A salesman must visit  $n$  towns successively and return to his point of departure in a minimum time. We denote by  $t_{ij}$  the time to travel between the town  $i$  and the town  $j$  (possibly different from  $t_{ji}$ ). We then draw the oriented graph of the  $n$  towns linked by arcs weighted by the  $(t_{ij})$  (see Figure 9.1). We must then find a cycle in this graph which passes once and once only through all the towns. It is possible to cast this problem as a linear programming problem in integer variables (see Section 11.6.1 for a solution to the method). •

**Example 9.1.5 (path of minimum cost)** Take an oriented graph  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ , where  $\mathcal{N}$  is the set of nodes and  $\mathcal{A} \subset \mathcal{N} \times \mathcal{N}$  is a set of arcs linking these nodes.

We associate with each arc  $(k, m) \in \mathcal{A}$  a cost  $w(k, m)$ . We fix a node as the origin  $i$  and a node as the destination  $j$ . The problem of the minimum cost path consists of finding a path of the graph going from  $i$  to  $j$  in an arbitrary number of steps which has total minimum cost. In other words, we look for a sequence of nodes  $i = \ell_0, \dots, \ell_T = j$  such that  $(\ell_r, \ell_{r+1}) \in \mathcal{A}$  for all  $r = 0, 1, \dots, T-1$  and  $w(\ell_0, \ell_1) + \dots + w(\ell_{T-1}, \ell_T)$  is minimal.

When the nodes are towns, so that  $\mathcal{A}$  is the set of direct routes from one town to the other, and  $w(k, m)$  is the distance between the towns  $k$  and  $m$ , we recover the classical problem of the shortest path. In spite of its combinatorial aspect, this problem is easy to solve even when it is large. •

Here is a very simple algebraic example which comes, for example, from the finite element discretization of the Stokes equations (see Section 6.3.4).

**Example 9.1.6 (quadratic optimization with linear constraints)** Let  $A$  be a square matrix of order  $n$ , which is symmetric and positive definite. Let  $B$  be a rectangular matrix of size  $m \times n$ . Let  $b$  be a vector of  $\mathbb{R}^m$ . We want to solve the problem

$$\inf_{x \in \text{Ker} B} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\}.$$

The constraint of belonging to  $\text{Ker} B$  makes this minimization nonobvious (see Section 10.2.2 for its solution). •

Another simple algebraic example is that of the Rayleigh quotient which allows us to calculate the eigenvalues and eigenvectors of a symmetric matrix.

**Example 9.1.7 (first eigenvalue)** Let  $A$  be a square matrix of order  $n$ , which is symmetric. We want to characterize and calculate the solutions of

$$\inf_{x \in \mathbb{R}^n, \|x\|=1} Ax \cdot x,$$

where  $\|x\|$  is the Euclidean norm of  $x$ . We see that they are the eigenvectors of  $A$  associated with its smallest eigenvalue (cf. Section 10.2.2). •

Let us consider a classical example in economics.

**Example 9.1.8 (household consumption)** We consider a household which consumes  $n$  types of goods whose prices form a vector  $p \in \mathbb{R}_+^n$ . The available income is a real number  $b > 0$ , and the choices for consumption are assumed to be modelled by a utility function  $u(x)$  from  $\mathbb{R}_+^n$  into  $\mathbb{R}$  (increasing and concave), which measures the benefit that the household gains from the consumption of the quantity  $x$  of  $n$  goods. The consumption of the household will be the vector  $x^*$  which will realize the maximum of

$$\max_{x \in \mathbb{R}_+^n, x \cdot p \leq b} u(x),$$

that is to say which maximizes the utility under a maximal budget constraint (see Section 10.3.2 for the solution). •

Let us pass to an example of optimization of a system modelled by an ordinary differential equation, that is to say to a problem of **optimal command**.

**Example 9.1.9 (optimal command)** We consider a linear differential system with quadratic criterion. The aim is to guide a robot (or a spacecraft, a vehicle, etc.) so that it follows a predefined trajectory ‘as closely as possible’. The state of the robot at the moment  $t$  is represented by a function  $y(t)$  with values in  $\mathbb{R}^N$  (typically, the position and the velocity). We act on the robot through a command  $v(t)$  with values in  $\mathbb{R}^M$  (typically, the engine power, the direction of the wheels, etc.). In the presence of forces  $f(t) \in \mathbb{R}^N$  the laws of mechanics lead to a system of ordinary differential equations (assumed linear for simplicity)

$$\begin{cases} \frac{dy}{dt} = Ay + Bv + f & \text{for } 0 \leq t \leq T \\ y(0) = y_0 \end{cases} \quad (9.1)$$

where  $y_0 \in \mathbb{R}^N$  is the initial state of the system,  $A$  and  $B$  are two constant matrices of respective dimensions  $N \times N$  and  $N \times M$ . We denote by  $z(t)$  a ‘target’ trajectory and  $z_T$  a final ‘target’ position. To approximate these targets as well as possible and to minimize the cost of the control, we introduce three symmetric positive matrices  $R, Q, D$  where only  $R$  is further assumed to be positive definite. We then define a quadratic criterion

$$J(v) = \int_0^T Rv(t) \cdot v(t) dt + \int_0^T Q(y - z)(t) \cdot (y - z)(t) dt + D(y(T) - z_T) \cdot (y(T) - z_T).$$

Let us remark that the function  $y(t)$  depends on the variable  $v$  through (9.1). As the admissible commands are possibly limited (the power of a motor is often bounded...), we introduce a convex closed nonempty set  $K$  of  $\mathbb{R}^M$  which represents the set of admissible commands. The problem is therefore to solve

$$\inf_{v(t) \in K, 0 \leq t \leq T} J(v).$$

It will be necessary, of course, to specify in which function spaces we minimize  $J(v)$  and we define the solution  $y$  of (9.1) (see Section 10.4.2 for the solution). •

**Example 9.1.10 (minimization of a mechanical energy)** We want to minimize the mechanical energy of a membrane or the electrostatic energy of a conductor. We refer to Chapters 1 and 5 for more detail on the modelling and the mathematical notation. Let  $\Omega$  be an open bounded set of  $\mathbb{R}^N$  and take  $f \in L^2(\Omega)$ . From proposition

5.2.7, to solve the Dirichlet problem for the Laplacian, we must minimize the energy  $J(v)$  defined for  $v \in H_0^1(\Omega)$  by

$$J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx.$$

In other words, we want to solve

$$\inf_{v \in H_0^1(\Omega)} J(v).$$

We can pose the same problem for more complicated equations than the Laplacian such as the Stokes equations. According to exercise 5.3.10, the solution of the Stokes equations is equivalent to the minimization

$$\inf_{v \in H_0^1(\Omega)^N \text{ such that } \operatorname{div} v = 0} \left\{ J(v) = \frac{\mu}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f \cdot v dx \right\}.$$

We remark that there is an ‘incompressibility’ constraint in this minimization, and that the pressure is absent from the energy. We see that these two facts are closely linked. •

Let us now give an example from the calculus of variations. Historically, this is one of the oldest problems in optimization, solved by Zénodore around two centuries before our era and whose complete proof is due to Weierstrass towards the end of the nineteenth century.

**Example 9.1.11 (the Didon problem)** Virgil recounts in the Aeneid that when the queen Didon founded the city of Carthage, all the land that was allocated was ‘as much land as could be enclosed by the skin of an ox’. She then cut this skin into fine strips and encircled the future city situated beside the sea. The question was therefore to find the greatest area possible bounded by a line (the shore) with given fixed boundary length. The answer is of course a half disc (see exercise 10.2.10). In slightly simplified mathematical terms, the problem is to find the plane curve of fixed length  $l \geq 0$  which encloses, with the segment linking its two ends, the maximum area. In other words, we solve

$$\sup \int_0^{\xi} y(x) dx,$$

with the constraints

$$\xi \geq 0, \quad y(0) = 0, \quad \int_0^{\xi} \sqrt{1 + y'(x)^2} dx = l,$$

where  $\xi$  is the end of the segment and  $y(x)$  the position of the curve above the point  $x$  of the segment. •

Let us now come to the optimization of a **distributed system**, that is to say one modelled by a partial differential equation.

**Example 9.1.12 (control of a membrane)** We consider an elastic membrane, fixed at its boundary, and deformed under the action of a force  $f$ . As we have seen in Section 1.3.3, this problem is modelled by

$$\begin{cases} -\Delta u = f + v & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where  $u$  is the vertical displacement of the membrane and  $v$  is a control force available to us. This control is typically a piezoelectric actuator which acts on a part  $\omega$  of the domain  $\Omega$  with a limited intensity. We therefore define the set of admissible controls

$$K = \{v(x) \text{ such that } v_{\min}(x) \leq v(x) \leq v_{\max}(x) \text{ in } \omega \text{ and } v = 0 \text{ in } \Omega \setminus \omega\},$$

where  $v_{\min}$  and  $v_{\max}$  are two given functions. We look for the control which gives the displacement  $u$  as close as possible to a desired displacement  $u_0$ , and which has a moderate cost. We therefore define a criterion

$$J(v) = \frac{1}{2} \int_{\Omega} (|u - u_0|^2 + c|v|^2) dx,$$

with  $c > 0$ . The control problem is written

$$\inf_{v \in K} J(v).$$

We must still, of course, specify the choice of function spaces for  $v$  and the other data of this problem (see Section 10.4.3 for the solution). •

### 9.1.3 Definitions and notation

Optimization has a particular vocabulary: let us introduce some classical notation and definitions. We consider principally some minimization problems (knowing that it is enough to change the sign to obtain a maximization problem).

First of all, the space in which the problem lies, denoted as  $V$ , is assumed to be a normed vector space, that is to say equipped with a norm denoted by  $\|v\|$ . In Section 9.1.4  $V$  will be the space  $\mathbb{R}^N$ , while in the following section  $V$  will be a real Hilbert space (we could equally well consider the more general case of a Banach space, that is to say a complete normed vector space). We also have a subset  $K \subset V$  where we will look for the solution: we say that  $K$  is the set of **admissible** elements of the problem, or that  $K$  defines the **constraints** imposed on the problem. Finally, the **criterion**, or the **cost function**, or the **objective function**, to be minimized, denoted by  $J$ , is a function defined over  $K$  with values in  $\mathbb{R}$ . The problem studied will therefore be denoted as

$$\inf_{v \in K \subset V} J(v). \quad (9.2)$$

When we use the notation  $\inf$  for a minimization problem, this indicates that we do not know *a priori*, if the minimum value is attained, that is to say if there exists  $\bar{v} \in K$  such that

$$J(\bar{v}) = \inf_{v \in K \subset V} J(v).$$

If we want to indicate that the minimum value is attained, we prefer the notation

$$\min_{v \in K \subset V} J(v),$$

but this is not a universal convention (though extremely widespread). For the maximization problems, the notation  $\sup$  and  $\max$  replace  $\inf$  and  $\min$ , respectively. Let us specify some basic definitions.

**Definition 9.1.1** *We say that  $u$  is a local minimum (or minimum point) of  $J$  over  $K$  if and only if*

$$u \in K \quad \text{and} \quad \exists \delta > 0, \quad \forall v \in K, \quad \|v - u\| < \delta \implies J(v) \geq J(u).$$

*We say that  $u$  is a global minimum (or minimum point) of  $J$  over  $K$  if and only if*

$$u \in K \quad \text{and} \quad J(v) \geq J(u) \quad \forall v \in K.$$

**Definition 9.1.2** *We say infimum of  $J$  over  $K$  (or, more usually, minimum value), and which we denote by (9.2), to mean the upper bound in  $\mathbb{R}$  of the constants which bound  $J$  below on  $K$ . If  $J$  is not bounded below  $K$ , then the infimum is  $-\infty$ . If  $K$  is empty, by convention, the infimum is  $+\infty$ .*

*A minimizing sequence of  $J$  in  $K$  is a sequence  $(u^n)_{n \in \mathbb{N}}$  such that*

$$u^n \in K \quad \forall n \quad \text{and} \quad \lim_{n \rightarrow +\infty} J(u^n) = \inf_{v \in K} J(v).$$

By the very definition of the infimum of  $J$  over  $K$  there always exist minimizing sequences.

### 9.1.4 Optimization in finite dimensions

Let us interest ourselves now in the question of the **existence of minima** for optimization problems posed in finite dimensions. We shall assume in this section (without loss of generality) that  $V = \mathbb{R}^N$  provided with the usual scalar product  $u \cdot v = \sum_{i=1}^N u_i v_i$  and with the Euclidean norm  $\|u\| = \sqrt{u \cdot u}$ .

A general result guaranteeing the existence of a minimum is the following.

**Theorem 9.1.3 (existence of a minimum in finite dimensions)** *Let  $K$  be a closed nonempty set of  $\mathbb{R}^N$ , and  $J$  a continuous function over  $K$  with values in  $\mathbb{R}$  satisfying the property, called ‘infinite at infinity’,*

$$\forall (u^n)_{n \geq 0} \text{ a sequence in } K, \quad \lim_{n \rightarrow +\infty} \|u^n\| = +\infty \implies \lim_{n \rightarrow +\infty} J(u^n) = +\infty. \quad (9.3)$$

Then there exists at least one minimum point of  $J$  over  $K$ . Further, from every minimizing sequence of  $J$  over  $K$  we can extract a subsequence converging to a minimum point over  $K$ .

**Proof.** Let  $(u^n)$  be a minimizing sequence of  $J$  over  $K$ . Condition (9.3) implies that  $u^n$  is bounded since  $J(u^n)$  is a sequence of bounded reals. Therefore, there exists a subsequence  $(u^{n_k})$  which converges to a point  $u$  of  $\mathbb{R}^N$ . But  $u \in K$  since  $K$  is closed, and  $J(u^{n_k})$  converges to  $J(u)$  by continuity, from which  $J(u) = \inf_{v \in K} J(v)$  according to definition 9.1.2.  $\square$

**Remark 9.1.4** Let us note that property (9.3), which assures that every minimizing sequence of  $J$  over  $K$  is bounded, is automatically satisfied if  $K$  is bounded. When the set  $K$  is not bounded, this condition means that, in  $K$ ,  $J$  is **infinite at infinity**.  $\bullet$

**Exercise 9.1.1** Show by example that the fact that  $K$  is closed or that  $J$  is continuous is in general necessary for the existence of a minimum. Give an example of a continuous function which is bounded below from  $\mathbb{R}$  into  $\mathbb{R}$  and which does not have a minimum over  $\mathbb{R}$ .

**Exercise 9.1.2** Show that we can replace the property 'infinite at infinity' (9.3) by the weaker condition

$$\inf_{v \in K} J(v) < \lim_{R \rightarrow +\infty} \left( \inf_{\|v\| \geq R} J(v) \right).$$

**Exercise 9.1.3** Show that we can replace the continuity of  $J$  by the lower semi-continuity of  $J$  defined by

$$\forall (u^n)_{n \geq 0} \text{ a sequence in } K, \quad \lim_{n \rightarrow +\infty} u^n = u \implies \liminf_{n \rightarrow +\infty} J(u^n) \geq J(u).$$

**Exercise 9.1.4** Show that there exists a minimum for examples 9.1.1, 9.1.6, and 9.1.7.

**Exercise 9.1.5** Let  $a$  and  $b$  be two real numbers with  $0 < a < b$ , and for  $n \in \mathbb{N}^*$ , let  $\mathcal{P}_n$  be the set of polynomials  $P$  of degree less than or equal to  $n$  such that  $P(0) = 1$ . For  $P \in \mathcal{P}_n$ , we denote  $\|P\| = \max_{x \in [a, b]} |P(x)|$ .

1. Show that the problem

$$\inf_{P \in \mathcal{P}_n} \|P\| \tag{9.4}$$

has a solution.

2. We recall that the Chebyshev polynomials  $T_n(X)$  are defined by the relations

$$T_0(X) = 1, \quad T_1(X) = X, \quad T_{n+1}(X) = 2XT_n(X) - T_{n-1}(X).$$

Show that the degree of  $T_n$  is equal to  $n$  and that for all  $\theta \in \mathbb{R}$ ,  $T_n(\cos \theta) = \cos(n\theta)$ . Deduce the existence of  $n+1$  real numbers  $\xi_0^n = 1 > \xi_1^n > \xi_2^n > \dots > \xi_n^n = -1$  such that  $T_n(\xi_k^n) = (-1)^k$  for  $0 \leq k \leq n$  and that  $\max_{-1 \leq x \leq 1} |T_n(x)| = 1$ .

3. Show that the unique solution of (9.4) is the polynomial

$$P(X) = \frac{1}{T_n((b+a)/(b-a))} T_n\left(\frac{(b+a)/2 - X}{(b-a)/2}\right).$$

## 9.2 Existence of a minimum in infinite dimensions

### 9.2.1 Examples of nonexistence

This section is dedicated to two examples showing that the existence of a minimum in infinite dimensions is **not absolutely guaranteed** by conditions like those used in the statement of theorem 9.1.3. This difficulty is closely linked to the fact that in infinite dimensions the closed bounded sets are not compact!

**Example 9.2.1** Take the Hilbert space (of infinite dimensions) of square summable sequences in  $\mathbb{R}$

$$\ell_2(\mathbb{R}) = \left\{ x = (x_i)_{i \geq 1} \text{ such that } \sum_{i=1}^{+\infty} x_i^2 < +\infty \right\},$$

equipped with the scalar product  $\langle x, y \rangle = \sum_{i=1}^{+\infty} x_i y_i$ . We consider the function  $J$  defined over  $\ell_2(\mathbb{R})$  by

$$J(x) = (\|x\|^2 - 1)^2 + \sum_{i=1}^{+\infty} \frac{x_i^2}{i}.$$

Taking  $K = \ell_2(\mathbb{R})$ , we consider the problem

$$\inf_{x \in \ell_2(\mathbb{R})} J(x), \tag{9.5}$$

for which we shall show that there does not exist a minimum point. We verify first of all that

$$\left( \inf_{x \in \ell_2(\mathbb{R})} J(x) \right) = 0.$$

Let us introduce the sequence  $x^n$  in  $\ell_2(\mathbb{R})$  defined by  $x_i^n = \delta_{in}$  for all  $i \geq 1$ . We verify easily that

$$J(x^n) = \frac{1}{n} \rightarrow 0 \quad \text{when } n \rightarrow +\infty.$$

As  $J$  is positive, we deduce that  $x^n$  is a minimizing sequence and that the minimum value is zero. However, it is evident that there does not exist any  $\bar{x} \in \ell_2(\mathbb{R})$  such that  $J(\bar{x}) = 0$ . Consequently, there does not exist a minimum point for (9.5). We see in this example that the minimizing sequence  $x^n$  is not compact in  $\ell_2(\mathbb{R})$  (although it is bounded). •



Here now is a model example which is not without similarity to the energy minimization problems that we have met in the solution of partial differential equations (see, for example, proposition 5.2.7). In spite of its simplified character, this example is very representative of realistic and practical problems in minimization of phase change energy in material science.

**Example 9.2.2** We consider the Sobolev space  $V = H^1(0, 1)$  equipped with the norm  $\|v\| = \left( \int_0^1 (v'(x)^2 + v(x)^2) dx \right)^{1/2}$  (which is an infinite dimensional Hilbert space, see Chapter 4). We set  $K = V$  and, for  $1 \geq h > 0$ , we consider

$$J_h(v) = \int_0^1 \left( (|v'(x)| - h)^2 + v(x)^2 \right) dx.$$

The mapping  $J$  is continuous over  $V$ , and the condition (9.3) is satisfied as

$$J_h(v) = \|v\|^2 - 2h \int_0^1 |v'(x)| dx + h^2 \geq \|v\|^2 - \frac{1}{2} \int_0^1 v'(x)^2 dx - h^2 \geq \frac{\|v\|^2}{2} - h^2.$$

Let us show that

$$\inf_{v \in V} J_h(v) = 0, \quad (9.6)$$

which will imply that there does not exist a minimum of  $J_h$  over  $V$ : in effect, if (9.6) holds and if  $u$  was a minimum of  $J_h$  over  $V$ , we should have  $J_h(u) = 0$ , from which  $u \equiv 0$  and  $|u'| \equiv h > 0$  (almost everywhere) over  $(0, 1)$ , which is impossible.

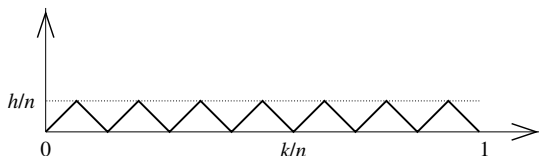


Figure 9.2. Minimizing sequence  $u^n$  for example 9.2.2.

To obtain (9.6), we construct a minimizing sequence  $(u^n)$  defined for  $n \geq 1$  by

$$u^n(x) = \begin{cases} h(x - \frac{k}{n}) & \text{if } \frac{k}{n} \leq x \leq \frac{2k+1}{2n}, \\ h(\frac{k+1}{n} - x) & \text{if } \frac{2k+1}{2n} \leq x \leq \frac{k+1}{n}, \end{cases} \quad \text{for } 0 \leq k \leq n-1,$$

as Figure 9.2 shows. We easily see that  $u^n \in V$  and the derivative  $(u^n)'(x)$  only takes two values:  $+h$  and  $-h$ . Consequently,  $J_h(u^n) = \int_0^1 u^n(x)^2 dx = h^2/4n$ , which proves (9.6), that is to say that  $J_h$  does not have a minimum point over  $V$ . And yet, if  $h = 0$ , it is clear that  $J_0$  has a unique minimum point  $v \equiv 0$ ! •

**Exercise 9.2.1** Modify the preceding construction to show that there are no more minima of  $J$  over  $C^1[0, 1]$ .

In the light of these counterexamples, let us examine the difficulty which occurs in infinite dimensions and under what hypotheses we can hope to obtain an existence result for a minimization problem posed in an infinite dimensional Hilbert space.

Let  $V$  be a vector space with norm  $\|v\|$ . Let  $J$  be a function defined on subset  $K$  of  $V$  with values in  $\mathbb{R}$ , satisfying the condition (9.3) (infinite at infinity). Then, every minimizing sequence  $(u^n)$  of the problem

$$\inf_{v \in K} J(v) \quad (9.7)$$

is bounded. In finite dimensions (if  $V = \mathbb{R}^N$ ), we finish easily as in Section 9.1.4 by using the compactness of closed bounded sets (and assuming that  $K$  is closed and that  $J$  is continuous or lower semicontinuous). Unfortunately, such a result is false in infinite dimensions, as we have just noted. Generally we can finish if the triplet  $(V, K, J)$  satisfies the following condition: for every sequence  $(u^n)_{n \geq 1}$  in  $K$  such that  $\sup_{n \in \mathbb{N}} \|u^n\| < +\infty$  we have

$$\lim_{n \rightarrow +\infty} J(u^n) = \ell < +\infty \implies \exists u \in K \text{ such that } J(u) \leq \ell. \quad (9.8)$$

Thus, under conditions (9.3) and (9.8), problem (9.7) has a solution.

Unfortunately, condition (9.8) is not useful since it is not verifiable in general! We can however verify it for a particular class of problems, which are very important in theory and in practice: **convex** minimization problems. As we see in Section 9.2.3, if  $V$  is a Hilbert space,  $K$  a **convex** closed set of  $V$ , and  $J$  a continuous **convex** function over  $K$ , then (9.8) holds and problem (9.7) has a solution. The motivations to introduce these conditions are, on the one hand, that the convexity hypotheses are often natural in many applications, and on the other hand, that it is a rare class of problems for which the theory is sufficiently general and complete. But this does not mean that these conditions are the only ones which ensure the existence of a minimum! Nevertheless, outside of the convex framework developed in the following sections, difficulties of the type that we have met in the preceding counterexamples can occur.

## 9.2.2 Convex analysis

In everything that follows, we shall assume that  $V$  is a Hilbert space equipped with a scalar product  $\langle u, v \rangle$  and with associated norm  $\|v\|$ . Let us recall that a set  $K$  is convex if it contains all the segments linking any two of its points (see definition 12.1.9). Let us give some properties of convex functions.

**Definition 9.2.1** We say that a function  $J$  defined over a nonempty convex set  $K \subset V$  with values in  $\mathbb{R}$  is convex over  $K$  if and only if

$$J(\theta u + (1 - \theta)v) \leq \theta J(u) + (1 - \theta)J(v) \quad \forall u, v \in K, \quad \forall \theta \in [0, 1]. \quad (9.9)$$

Further,  $J$  is called *strictly convex* if the inequality (9.9) is strict when  $u \neq v$  and  $\theta \in ]0, 1[$ .

**Remark 9.2.2** If  $J$  is a mapping defined over  $K$  with values in  $\mathbb{R}$ , we call the **epi-graph** of  $J$  the set  $\text{Epi}(J) = \{(\lambda, v) \in \mathbb{R} \times K, \lambda \geq J(v)\}$ . Then  $J$  is convex if and only if  $\text{Epi}(J)$  is a convex set of  $\mathbb{R} \times V$ . •

**Exercise 9.2.2** Let  $J_1$  and  $J_2$  be two convex functions over  $V$ ,  $\lambda > 0$ , and  $\varphi$  a convex increasing function on an interval of  $\mathbb{R}$  containing the set  $J_1(V)$ . Show that  $J_1 + J_2$ ,  $\max(J_1, J_2)$ ,  $\lambda J_1$ , and  $\varphi \circ J_1$  are convex.

**Exercise 9.2.3** Let  $(L_i)_{i \in I}$  be a family (possibly infinite) of affine functions over  $V$ . Show that  $\sup_{i \in I} L_i$  is convex on  $V$ . Conversely, let  $J$  be a continuous convex function on  $V$ . Show that  $J$  is equal to  $\sup_{L_i \leq J} L_i$  where the functions  $L_i$  are affine.

For convex functions there is no difference between local and global minima as the following elementary result shows.

**Proposition 9.2.3** *If  $J$  is a convex function on a convex set  $K$ , every local minimum point of  $J$  over  $K$  is a global minimum and the set of minimum points is a convex set (possibly empty).*

*If further  $J$  is strictly convex, then there exists at most one minimum point.*

**Proof.** Let  $u$  be a local minimum of  $J$  over  $K$ . From definition 9.1.1, we can write

$$\exists \delta > 0, \forall w \in K, \|w - u\| < \delta \implies J(w) \geq J(u). \quad (9.10)$$

Take  $v \in K$ . For  $\theta \in ]0, 1[$  sufficiently small,  $w_\theta = \theta v + (1 - \theta)u$  satisfies  $\|w_\theta - u\| < \delta$  and  $w_\theta \in K$  as  $K$  is convex. Therefore,  $J(w_\theta) \geq J(u)$  according to (9.10), and the convexity of  $J$  implies that  $J(u) \leq J(w_\theta) \leq \theta J(v) + (1 - \theta)J(u)$ , which shows that  $J(u) \leq J(v)$ , that is to say that  $u$  is a global minimum over  $K$ .

On the other hand, if  $u_1$  and  $u_2$  are two minima and if  $\theta \in [0, 1]$ , then  $w = \theta u_1 + (1 - \theta)u_2$  is a minimum as  $w \in K$  and

$$\inf_{v \in K} J(v) \leq J(w) \leq \theta J(u_1) + (1 - \theta)J(u_2) = \inf_{v \in K} J(v).$$

The same argument with  $\theta \in ]0, 1[$  shows that, if  $J$  is strictly convex, then necessarily  $u_1 = u_2$ .  $\square$

In what follows we shall use the idea of ‘strong convexity’ which is **more restrictive** than strict convexity.

**Definition 9.2.4** *We say that a function  $J$  defined over a convex set  $K$  is strongly convex if and only if there exists  $\alpha > 0$  such that*

$$J\left(\frac{u+v}{2}\right) \leq \frac{J(u) + J(v)}{2} - \frac{\alpha}{8} \|u - v\|^2. \quad (9.11)$$

*We say also in this case that  $J$  is  $\alpha$ -convex.*

In definition 9.2.4, the strong convexity of  $J$  is only tested for convex combinations of weight  $\theta = 1/2$ . This is not a restriction for continuous functions as the following exercise shows.

**Exercise 9.2.4** If  $J$  is continuous and  $\alpha$ -convex, show that, for all  $\theta \in [0, 1]$ ,

$$J(\theta u + (1 - \theta)v) \leq \theta J(u) + (1 - \theta)J(v) - \frac{\alpha\theta(1 - \theta)}{2} \|u - v\|^2. \quad (9.12)$$

**Exercise 9.2.5** Let  $A$  be a symmetric matrix of order  $N$  and take  $b \in \mathbb{R}^N$ . For  $x \in \mathbb{R}^N$ , we set  $J(x) = \frac{1}{2}Ax \cdot x - b \cdot x$ . Show that  $J$  is convex if and only if  $A$  is positive semidefinite, and that  $J$  is strictly convex if and only if  $A$  is positive definite. In this last case, show that  $J$  is also strongly convex and find the best constant  $\alpha$ .

**Exercise 9.2.6** Let  $\Omega$  be an open set of  $\mathbb{R}^N$  and  $H^1(\Omega)$  the associated Sobolev space (see definition 4.3.1). Take the function  $J$  defined over  $\Omega$  by

$$J(v) = \frac{1}{2} \int_{\Omega} (|\nabla v(x)|^2 + v(x)^2) dx - \int_{\Omega} f(x)v(x) dx,$$

with  $f \in L^2(\Omega)$ . Show that  $J$  is strongly convex over  $H^1(\Omega)$ .

The following result will be essential to obtain an existence result for a minimum in infinite dimensions. In particular, it allows us to conclude that a function  $J$  which is strongly convex and continuous over a convex closed nonempty set  $K$  is ‘infinite at infinity’ in  $K$ , that is to say satisfies the property (9.3).

**Proposition 9.2.5** *If  $J$  is convex and continuous over a convex closed nonempty set  $K$ , then there exists a linear continuous form  $L \in V'$  and a constant  $\delta \in \mathbb{R}$  such that*

$$J(v) \geq L(v) + \delta \quad \forall v \in K. \quad (9.13)$$

*If further  $J$  is strongly convex over  $K$ , then there exist two constants  $\gamma > 0$  and  $\delta \in \mathbb{R}$  such that*

$$J(v) \geq \gamma \|v\|^2 - \delta \quad \forall v \in K. \quad (9.14)$$

**Proof.** Let us prove (9.13) first of all. If  $J$  is convex and continuous (or simply lower semicontinuous) over a convex closed nonempty set  $K$ , then its epigraph  $Epi(J)$  (defined in remark 9.2.2) is a convex closed nonempty set. Take  $v_0 \in K$  and  $\lambda_0 < J(v_0)$ . As  $(\lambda_0, v_0) \notin Epi(J)$ , we deduce from theorem 12.1.19, of the separation of a point and of a convex set, the existence of  $\alpha, \beta \in \mathbb{R}$  and of a continuous linear form  $L \in V'$  such that

$$\beta\lambda + L(v) > \alpha > \beta\lambda_0 + L(v_0) \quad \forall (\lambda, v) \in Epi(J). \quad (9.15)$$

As, for  $v$  fixed, we can take  $\lambda$  arbitrarily large in the left-hand side of (9.15), it is clear that  $\beta \geq 0$ ; further, as we can take  $v = v_0$  in the left-hand side of (9.15),  $\beta$  cannot be

zero. We therefore have  $\beta > 0$  and we deduce from (9.15) that  $J(v) + L(v)/\beta > \alpha/\beta$  for all  $v \in K$ , which proves (9.13).

Let us now prove (9.14). Again take  $v_0 \in K$  fixed. For every  $v \in K$ , (9.11) and (9.13) imply that

$$\frac{J(v)}{2} + \frac{J(v_0)}{2} \geq J\left(\frac{v+v_0}{2}\right) + \frac{\alpha}{8}\|v-v_0\|^2 \geq \frac{L(v)+L(v_0)}{2} + \frac{\alpha}{8}\|v-v_0\|^2 + \delta.$$

We deduce

$$J(v) \geq \frac{\alpha}{4}\|v\|^2 - \frac{\alpha}{2}\langle v, v_0 \rangle + L(v) + C_1,$$

with  $C_1 = (\alpha/4)\|v_0\|^2 + L(v_0) - J(v_0) + 2\delta$ . From the Cauchy–Schwarz inequality applied to  $\langle v, v_0 \rangle$  and the continuity of  $L$ , that is,  $|L(v)| \leq \|L\|_{V'}\|v\|$  (see definition 12.1.17), we have

$$J(v) \geq \frac{\alpha}{4}\|v\|^2 - \left(\|L\|_{V'} + \frac{\alpha\|v_0\|}{2}\right)\|v\| + C_1 \geq \frac{\alpha}{8}\|v\|^2 - C,$$

for  $C \in \mathbb{R}$  well chosen. □

Let us finish this section with an agreeable property of convex functions: ‘proper’ convex functions (that is to say which do not take the value  $+\infty$ ) are continuous.

**Exercise 9.2.7** Take  $v_0 \in V$  and let  $J$  be a convex function bounded above over an open ball of centre  $v_0$ . Show that  $J$  is bounded below and continuous over this ball.

### 9.2.3 Existence results

We can now state a first existence result for the minimum in the particular case where  $J$  is strongly convex ( $\alpha$ -convex).

**Theorem 9.2.6 (existence of a minimum, strongly convex case)** *Let  $K$  be a convex closed nonempty set of a Hilbert  $V$  and  $J$  an  $\alpha$ -convex continuous function over  $K$ . Then, there exists a unique minimum  $u$  of  $J$  over  $K$  and we have*

$$\|v-u\|^2 \leq \frac{4}{\alpha}[J(v) - J(u)] \quad \forall v \in K. \quad (9.16)$$

*In particular, every minimizing sequence of  $J$  over the set  $K$  converges to  $u$ .*

**Proof.** Let  $(u^n)$  be a minimizing sequence of  $J$  over  $K$ . From (9.14),  $J$  is bounded below over  $K$  and, for  $n, m \in \mathbb{N}$  the property (9.11) of strong convexity implies that

$$\begin{aligned} & \frac{\alpha}{8}\|u^n - u^m\|^2 + J\left(\frac{u^n + u^m}{2}\right) - \inf_{v \in K} J(v) \\ & \leq \frac{1}{2}\left(J(u^n) - \inf_{v \in K} J(v)\right) + \frac{1}{2}\left(J(u^m) - \inf_{v \in K} J(v)\right), \end{aligned}$$

which shows that the sequence  $(u^n)$  is Cauchy, and therefore converges to a limit  $u$ , which is necessarily a minimum of  $J$  over  $K$  as  $J$  is continuous and  $K$  closed. The uniqueness of the minimum point has been shown in proposition 9.2.3. Finally, if  $v \in K$ ,  $(u+v)/2 \in K$  since  $K$  is convex, from which, thanks to (9.11),

$$\frac{\alpha}{8} \|u - v\|^2 \leq \frac{J(u)}{2} + \frac{J(v)}{2} - J\left(\frac{u+v}{2}\right) \leq \frac{J(v) - J(u)}{2},$$

since  $J\left(\frac{u+v}{2}\right) \geq J(u)$ . □

**Exercise 9.2.8** Show that theorem 9.2.6 applies to example 9.1.10 (use the Poincaré inequality in  $H_0^1(\Omega)$ ).

**Exercise 9.2.9** Generalize exercise 9.2.8 to the different models met in Chapter 5: Laplacian with Neumann boundary conditions (see proposition 5.2.16), elasticity (see exercise 5.3.3), Stokes (see exercise 5.3.10).

It is possible to generalize theorem 9.2.6 to the case of functions  $J$  which are only convex (and not strongly convex). However, while much of the proof of theorem 9.2.6 is elementary, much of the following theorem is delicate. It relies in particular on the idea of weak convergence that we can consider as ‘supplementary’ in the framework of this course.

**Theorem 9.2.7 (existence of a minimum, convex case)** *Let  $K$  be a convex closed nonempty set of a Hilbert space  $V$ , and  $J$  a continuous convex function over  $K$ , which is ‘infinite at infinity’ in  $K$ , that is to say which satisfies condition (9.3), that is,*

$$\forall (u^n)_{n \geq 0} \text{ a sequence in } K, \quad \lim_{n \rightarrow +\infty} \|u^n\| = +\infty \implies \lim_{n \rightarrow +\infty} J(u^n) = +\infty.$$

*Then there exists a minimum of  $J$  over  $K$ .*

**Remark 9.2.8** Theorem 9.2.7 gives the existence of a minimum like the preceding theorem 9.2.6, but says nothing about the uniqueness or the error estimate (9.16). We remark in passing that (9.16) will be very useful for the study of numerical minimization algorithms since it gives an estimate of the rate of convergence of a minimizing sequence  $(u^n)$  to the minimum point  $u$ . •

**Remark 9.2.9** The theorem 9.2.7 remains true if we simply assume that  $V$  is a reflexive Banach space (that is, that the dual of  $V'$  is  $V$ ). •

We indicate briefly how we can prove theorem 9.2.7 in the case of a separable Hilbert space (that is to say which has a countable Hilbertian basis, see proposition 12.1.15). We define the idea of **weak convergence** in  $V$ .

**Definition 9.2.10** We say that a sequence  $(u^n)$  of  $V$  converges weakly to  $u \in V$  if

$$\forall v \in V, \quad \lim_{n \rightarrow +\infty} \langle u^n, v \rangle = \langle u, v \rangle.$$

Let  $(e^i)_{i \geq 1}$  be a Hilbertian basis of  $V$ . If we denote by  $u_i^n = \langle u^n, e^i \rangle$  the components in this basis of a sequence  $u^n$ , which is uniformly bounded in  $V$ , it is easy to verify that the definition 9.2.10 of the weak convergence is equivalent to **convergence of all the sequences of components**  $(u_i^n)_{n \geq 1}$  for  $i \geq 1$ .

As its name indicates, weak convergence is a ‘weaker’ idea than the usual convergence in  $V$ , as  $\lim_{n \rightarrow +\infty} \|u^n - u\| = 0$  implies that  $u^n$  converges weakly to  $u$ . Conversely, in infinite dimensions there exist sequences which converge weakly but not in the usual sense (which we sometimes call ‘strong convergence’ in contrast). For example, the sequence  $u^n = e^n$  converges weakly to zero, but not strongly since it has constant norm equal to 1. The interest in weak convergence comes from the following result.

**Lemma 9.2.11** From every sequence  $u^n$  bounded in  $V$  we can extract a subsequence which converges weakly.

**Proof.** As the sequence  $u^n$  is bounded, each sequence of a component  $u_i^n$  is bounded in  $\mathbb{R}$ . For each  $i$ , there therefore exists a subsequence, denoted  $u_i^{n_i}$ , which converges to a limit  $u_i$ . By a process of diagonal extraction of sequences we then obtain a common subsequence  $n'$  such that, for all  $i$ ,  $u_i^{n'}$  converges to  $u_i$ . This proves that  $u^{n'}$  converges weakly to  $u$  (we verify that  $u \in V$ ).  $\square$

If we say ‘closed half-space’ of  $V$  to mean every set of the form  $\{v \in V, L(v) \leq \alpha\}$ , where  $L$  is a continuous linear form not identically zero over  $V$  and  $\alpha \in \mathbb{R}$ , we can conveniently characterize closed convex sets.

**Lemma 9.2.12** A closed convex set  $K$  of  $V$  is the intersection of the closed half-spaces which contain  $K$ .

**Proof.** It is clear that  $K$  is included in the intersection of the closed half-spaces which contain it. Conversely, assume that there exists a point  $u_0$  of this intersection which does not belong to  $K$ . We can then apply the theorem 12.1.19 of separation of a point and of a convex set and therefore construct a closed half-space which contains  $K$  but not  $u_0$ . This is a contradiction with the definition of  $u_0$ , therefore  $u_0 \in K$ .  $\square$

**Lemma 9.2.13** Let  $K$  be a convex closed nonempty set of  $V$ . Then  $K$  is closed for weak convergence.

Further, if  $J$  is convex and lower semicontinuous over  $K$  (see exercise 9.1.3 for this idea), then  $J$  is also lower semicontinuous over  $K$  for weak convergence.

**Proof.** By definition, if  $u^n$  converges weakly to  $u$ , then  $L(u^n)$  converges to  $L(u)$ . Consequently, a closed half-space of  $V$  is closed for weak convergence. Lemma 9.2.12 allows us to reach the same conclusion for  $K$ .

From the hypotheses on  $J$ , the set  $\text{Epi}(J)$  (defined in remark 9.2.2) is a convex closed set of  $\mathbb{R} \times V$ , therefore it is also closed for weak convergence. We then easily deduce the following result: if the sequence  $(v^n)$  tends weakly to  $v$  in  $K$ , then  $\liminf_{n \rightarrow +\infty} J(v^n) \geq J(v)$ .  $\square$

We now have all the ingredients to finish.

**Proof of theorem 9.2.7.** From (9.3), every minimizing sequence  $(u^n)$  is bounded. We then deduce from lemma 9.2.11 that there exists a subsequence  $(u^{n'})$  converging weakly to a limit  $u \in V$ . But, according to lemma 9.2.13,  $u \in K$  and

$$J(u) \leq \liminf_k J(u^{n_k}) = \inf_{v \in K} J(v).$$

The point  $u$  is therefore a minimum of  $u$  over  $K$ . □



*This page intentionally left blank*

# 10 Optimality conditions and algorithms

---

## 10.1 Generalities

### 10.1.1 Introduction

In Chapter 9 we were interested in questions of existence of a minimum in optimization problems. In this chapter, we shall obtain necessary and sometimes sufficient conditions for minimality. The objective is in a certain way much more practical, since these optimality conditions will be more often used to try to **calculate a minimum** (sometimes even without having shown its existence!). The general idea of optimality conditions is the same as writing that the derivative must be zero, when we calculate the extremum of a function over  $\mathbb{R}$ .

These conditions will therefore be expressed with the help of the first derivative (conditions of order 1) or second derivative (conditions of order 2). Above all we will obtain **necessary** conditions for optimality, but the use of the second derivative or the introduction of convexity hypotheses will also allow us to obtain **sufficient** conditions, and to distinguish between minima and maxima.

These optimality conditions generalize the following elementary remark: if  $x_0$  is a local minimum point of  $J$  on the interval  $[a, b] \subset \mathbb{R}$  ( $J$  being a differentiable function on  $[a, b]$ ), then we have

$$J'(x_0) \geq 0 \quad \text{if } x_0 = a, \quad J'(x_0) = 0 \quad \text{if } x_0 \in ]a, b[, \quad J'(x_0) \leq 0 \quad \text{if } x_0 = b.$$

Even if it is well known to the reader, let us recall the proof of this remark: if  $x_0 \in [a, b[$ , we can choose  $x = x_0 + h$  with a small positive  $h > 0$ , and write  $J(x) \geq J(x_0)$ , from which  $J(x_0) + hJ'(x_0) + o(h) > J(x_0)$ , which gives  $J'(x_0) \geq 0$  dividing by  $h$  and letting  $h$  tend to 0. Likewise we obtain  $J'(x_0) \leq 0$  if  $x_0 \in ]a, b]$  by considering  $x = x_0 - h$ . Let us also remark (this is the second order condition) that if  $x_0 \in ]a, b[$

and if  $J'$  is differentiable at  $x_0$ , we then have  $J''(x_0) \geq 0$  (in effect, we have  $J(x_0) + (h^2/2)J''(x_0) + o(h^2) \geq J(x_0)$  for  $h$  small enough).

The strategy to obtain and to prove the minimality conditions is therefore clear: we take account of constraints ( $x \in [a, b]$  in the example above) to test the minimality of  $x_0$  in particular directions which respect the constraints ( $x_0 + h$  with  $h > 0$  if  $x_0 \in [a, b]$ ,  $x_0 - h$  with  $h > 0$  if  $x_0 \in ]a, b[$ ): we shall talk of **admissible directions**. We then use the definition of the derivative (and the second order Taylor formulas) to conclude. This is exactly what we shall do in what follows!

The plan of this chapter is the following. The remainder of this section is dedicated to making some notation precise and to recalling the elementary ideas of differentiability. Section 10.2 gives the form of the necessary optimality conditions in two essential cases: when the set of constraints is convex we obtain a **Euler inequality**; when it involves equality or inequality constraints, we obtain an equation using **Lagrange multipliers**. Section 10.3 is dedicated to the **Kuhn–Tucker theorem** which says that, under certain convexity hypotheses, the necessary conditions for optimality are also sufficient. We also give a brief outline of the theory of **duality**. Section 10.4 explores three applications of optimization to systems modelled by ordinary or partial differential equations. Finally, Section 10.5 treats **numerical optimization algorithms**. We principally study the **gradient** algorithms which are the most important in practice.

### 10.1.2 Differentiability

From now on (and we shall not systematically point this out any more), we assume that  $V$  is a real Hilbert space, and that  $J$  is a continuous function with values in  $\mathbb{R}$ . The scalar product in  $V$  is always denoted  $\langle u, v \rangle$  and the associated norm  $\|u\|$ .

Let us start by introducing the idea of a first derivative of  $J$  since we shall need this to write optimality conditions. When there are several variables (that is to say if the space  $V$  is not  $\mathbb{R}$ ), the ‘good’ theoretical idea of differentiability, called differentiability in the sense of Fréchet, is given by the following definition.

**Definition 10.1.1** *We say that the function  $J$ , defined on a neighbourhood of  $u \in V$  with values in  $\mathbb{R}$ , is differentiable in the sense of Fréchet at  $u$  if there exists a continuous linear form on  $V$ ,  $L \in V'$ , such that*

$$J(u + w) = J(u) + L(w) + o(w), \quad \text{with} \quad \lim_{w \rightarrow 0} \frac{|o(w)|}{\|w\|} = 0. \quad (10.1)$$

*We call  $L$  the derivative (or the differential, or the gradient) of  $J$  at  $u$  and we denote  $L = J'(u)$ .*

**Remark 10.1.2** Definition 10.1.1 is in fact valid if  $V$  is only a Banach space (we do not use the scalar product in (10.1)). However, if  $V$  is a Hilbert space, we can specify the relation (10.1) by identifying  $V$  and its dual  $V'$  thanks to the Riesz representation

theorem 12.1.18. In effect, there exists a unique  $p \in V$  such that  $\langle p, w \rangle = L(w)$ , therefore (10.1) becomes

$$J(u + w) = J(u) + \langle p, w \rangle + o(w), \quad \text{with } \lim_{w \rightarrow 0} \frac{|o(w)|}{\|w\|} = 0. \quad (10.2)$$

We also sometimes write  $p = J'(u)$ , which can lead to confusion. Formula (10.2) is often more ‘natural’ than (10.1), notably if  $V = \mathbb{R}^n$  or  $V = L^2(\Omega)$ . Conversely, there may be a slightly more delicate interpretation for more ‘complicated’ Hilbert spaces such as  $V = H^1(\Omega)$  (see exercise 10.1.5). •

In most applications, it is often sufficient to determine the continuous linear form  $L = J'(u) \in V'$  as we do not need the explicit expression of  $p = J'(u) \in V$  when  $V'$  is identified with  $V$ . In practice, it is easier to find the explicit expression for  $L$  than that for  $p$ , as the following exercises show.

**Exercise 10.1.1** Show that (10.1) implies the continuity of  $J$  at  $u$ . Show also that, if  $L_1, L_2$  satisfies

$$\begin{cases} J(u + w) \geq J(u) + L_1(w) + o(w), \\ J(u + w) \leq J(u) + L_2(w) + o(w), \end{cases} \quad (10.3)$$

then  $J$  is differentiable and  $L_1 = L_2 = J'(u)$ .

**Exercise 10.1.2 (essential!)** Let  $a$  be a continuous symmetric bilinear form over  $V \times V$ . Let  $L$  be a continuous linear form over  $V$ . We set  $J(u) = \frac{1}{2}a(u, u) - L(u)$ . Show that  $J$  is differentiable over  $V$  and that  $\langle J'(u), w \rangle = a(u, w) - L(w)$  for all  $u, w \in V$ .

**Exercise 10.1.3** Let  $A$  be an  $N \times N$  symmetric matrix and  $b \in \mathbb{R}^N$ . For  $x \in \mathbb{R}^N$ , we set  $J(x) = \frac{1}{2}Ax \cdot x - b \cdot x$ . Show that  $J$  is differentiable and that  $J'(x) = Ax - b$  for all  $x \in \mathbb{R}^N$ .

**Exercise 10.1.4** We return to exercise 10.1.2 with  $V = L^2(\Omega)$  ( $\Omega$  being an open set of  $\mathbb{R}^N$ ),  $a(u, v) = \int_{\Omega} uv \, dx$ , and  $L(u) = \int_{\Omega} fu \, dx$  with  $f \in L^2(\Omega)$ . By identifying  $V$  and  $V'$ , show that  $J'(u) = u - f$ .

**Exercise 10.1.5** We return to exercise 10.1.2 with  $V = H_0^1(\Omega)$  ( $\Omega$  being an open set of  $\mathbb{R}^N$ ) that we equip with the scalar product

$$\langle u, v \rangle = \int_{\Omega} (\nabla u \cdot \nabla v + uv) \, dx.$$

We set  $a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx$ , and  $L(u) = \int_{\Omega} fu \, dx$  with  $f \in L^2(\Omega)$ . Show (at least formally) that  $J'(u) = -\Delta u - f$  in  $V' = H^{-1}(\Omega)$ . Show that, if we identify  $V$  and  $V'$ , then  $J'(u) = u_0$  where  $u_0$  is the unique solution in  $H_0^1(\Omega)$  of

$$\begin{cases} -\Delta u_0 + u_0 = -\Delta u - f & \text{in } \Omega \\ u_0 = 0 & \text{on } \partial\Omega \end{cases}$$

**Exercise 10.1.6** Let  $\Omega$  be a bounded open set of  $\mathbb{R}^N$  (we can restrict ourselves to the case where  $N = 1$  with  $\Omega = ]0, 1[$ ). Let  $L = L(p, t, x)$  be a continuous function over  $\mathbb{R}^N \times \mathbb{R} \times \overline{\Omega}$ , differentiable with respect to  $p$  and  $t$  on this set, with partial derivatives  $\frac{\partial L}{\partial p}$  and  $\frac{\partial L}{\partial t}$  which are Lipschitzian over this set. We set  $V = H_0^1(\Omega)$  and  $J(v) = \int_{\Omega} L(\nabla v(x), v(x), x) dx$ .

1. Show that  $J$  is differentiable over  $H_0^1(\Omega)$  and that

$$\langle J'(u), w \rangle = \int_{\Omega} \left( \frac{\partial L}{\partial p}(\nabla u(x), u(x), x) \cdot \nabla w(x) + \frac{\partial L}{\partial t}(\nabla u(x), u(x), x) w(x) \right) dx.$$

2. If  $N = 1$  and  $\Omega = ]0, 1[$ , show that, if  $u \in H_0^1(0, 1)$  satisfies  $J'(u) = 0$ , then  $u$  satisfies

$$\frac{d}{dx} \left( \frac{\partial L}{\partial p}(u'(x), u(x), x) \right) - \frac{\partial L}{\partial t}(u'(x), u(x), x) = 0, \quad (10.4)$$

almost everywhere in the interval  $]0, 1[$ .

3. If  $L$  does not depend on  $x$  (that is,  $L = L(p, t)$ ) and if  $u \in C^2(]0, 1[)$  is a solution of the differential equation (10.4), show that the quantity

$$L(u'(x), u(x)) - u'(x) \frac{\partial L}{\partial p}(u'(x), u(x))$$

is constant on the interval  $[0, 1]$ .

**Remark 10.1.3** There exist other ideas of differentiability, weaker than in the sense of Fréchet. For example, we often meet the following definition. We say that the function  $J$ , defined on a neighbourhood of  $u \in V$  with values in  $\mathbb{R}$ , is differentiable in the sense of Gâteaux at  $u$  if there exists  $L \in V'$  such that

$$\forall w \in V, \quad \lim_{\delta \searrow 0+} \frac{J(u + \delta w) - J(u)}{\delta} = L(w). \quad (10.5)$$

The interest in this idea is that the verification of (10.5) is easier than that of (10.1). However, if a function is differentiable in the sense of Fréchet then it is also in the sense of Gâteaux, the converse is false, even in finite dimensions, as the following example shows in  $\mathbb{R}^2$

$$J(x, y) = \frac{x^6}{(y - x^2)^2 + x^8} \quad \text{for } (x, y) \neq (0, 0), \quad J(0, 0) = 0.$$

We shall, in what follows, say that a function is differentiable when it is in the sense of Fréchet, unless we explicitly mention otherwise. •

Let us now examine the basic properties of convex differentiable functions.

**Proposition 10.1.4** *Let  $J$  be a differentiable mapping from  $V$  into  $\mathbb{R}$ . The following assertions are equivalent*

$$J \text{ is convex over } V, \quad (10.6)$$

$$J(v) \geq J(u) + \langle J'(u), v - u \rangle \quad \forall u, v \in V, \quad (10.7)$$

$$\langle J'(u) - J'(v), u - v \rangle \geq 0 \quad \forall u, v \in V. \quad (10.8)$$

**Proposition 10.1.5** *Let  $J$  be a differentiable mapping from  $V$  into  $\mathbb{R}$  and  $\alpha > 0$ . The assertions following are equivalent*

$$J \text{ is } \alpha\text{-convex over } V, \quad (10.9)$$

$$J(v) \geq J(u) + \langle J'(u), v - u \rangle + \frac{\alpha}{2} \|v - u\|^2 \quad \forall u, v \in V, \quad (10.10)$$

$$\langle J'(u) - J'(v), u - v \rangle \geq \alpha \|u - v\|^2 \quad \forall u, v \in V. \quad (10.11)$$

**Remark 10.1.6** Conditions (10.10) and (10.7) have a simple geometrical interpretation. They signify that the convex function  $J(v)$  is always above its tangent plane at  $u$  (considered as an affine function of  $v$ ). Conditions (10.11) and (10.8) are coercivity hypotheses. In particular, if  $J(u) = \frac{1}{2}a(u, u) - L(u)$  with  $a$  a symmetric continuous bilinear form over  $V$  and  $L$  a continuous linear form over  $V$ , then exercise 10.1.2 shows that (10.11) is exactly the definition of the coercivity of  $a$ . •

**Proof.** It is enough to prove proposition 10.1.5 by observing that the case  $\alpha = 0$  gives proposition 10.1.4. Let us show that (10.9) implies (10.10). As  $J$  is  $\alpha$ -convex, we easily see (by recurrence) that, for all  $k \geq 1$ ,

$$J\left(\left(1 - \frac{1}{2^k}\right)u + \frac{1}{2^k}v\right) \leq \left(1 - \frac{1}{2^k}\right)J(u) + \frac{1}{2^k}J(v) - \frac{\alpha}{2^{k+1}}\left(1 - \frac{1}{2^k}\right)\|u - v\|^2,$$

from which

$$2^k \left[ J\left(u + \frac{1}{2^k}(v - u)\right) - J(u) \right] \leq J(v) - J(u) - \frac{\alpha}{2} \left(1 - \frac{1}{2^k}\right) \|u - v\|^2.$$

By letting  $k$  tend to  $+\infty$ , we find (10.10). To obtain (10.11) it is enough to add (10.10) to itself swapping  $u$  and  $v$ .

Let us show that (10.11) implies (10.9). For  $u, v \in V$  and  $t \in \mathbb{R}$ , we set  $\varphi(t) = J(u + t(v - u))$ . Then  $\varphi$  is differentiable and therefore continuous on  $\mathbb{R}$ , and  $\varphi'(t) = \langle J'(u + t(v - u)), v - u \rangle$ , so that, from (10.11)

$$\varphi'(t) - \varphi'(s) \geq \alpha(t - s)\|v - u\|^2 \quad \text{if } t \geq s. \quad (10.12)$$

Take  $\theta \in ]0, 1[$ . By integrating the inequality (10.12) from  $t = \theta$  to  $t = 1$  and from  $s = 0$  to  $s = \theta$ , we obtain

$$\theta\varphi(1) + (1 - \theta)\varphi(0) - \varphi(\theta) \geq \frac{\alpha\theta(1 - \theta)}{2} \|v - u\|^2,$$

that is to say (10.9). □

**Exercise 10.1.7** Show that a function  $J$  which is differentiable over  $V$  is strictly convex if and only if

$$J(v) > J(u) + \langle J'(u), v - u \rangle \quad \forall u, v \in V \text{ with } u \neq v,$$

or

$$\langle J'(u) - J'(v), u - v \rangle > 0 \quad \forall u, v \in V \text{ with } u \neq v.$$

Let us finish this section by defining the **second derivative** of  $J$ . Let us remark first of all that it is very easy to generalize the definition 10.1.1 of differentiability to the case of a function  $f$  defined over  $V$  with values in another Hilbert space  $W$  (and not only in  $\mathbb{R}$ ). We shall say that  $f$  is differentiable (in the sense of Fréchet) at  $u$  if there exists a continuous linear mapping  $L$  from  $V$  into  $W$  such that

$$f(u + w) = f(u) + L(w) + o(w), \quad \text{with } \lim_{w \rightarrow 0} \frac{\|o(w)\|}{\|w\|} = 0. \quad (10.13)$$

We call  $L = f'(u)$  the differential of  $f$  at  $u$ . Definition (10.13) is useful to define the derivative of  $f(u) = J'(u)$  which is a mapping from  $V$  into its dual  $V'$ .

**Definition 10.1.7** Let  $J$  be a function from  $V$  into  $\mathbb{R}$ . We say that  $J$  is twice differentiable at  $u \in V$  if  $J$  is differentiable in a neighbourhood of  $u$  and if its derivative  $J'(u)$  is differentiable at  $u$ . We denote by  $J''(u)$  the second derivative of  $J$  at  $u$  which satisfies

$$J'(u + w) = J'(u) + J''(u)w + o(w) \quad \text{with } \lim_{w \rightarrow 0} \frac{\|o(w)\|}{\|w\|} = 0.$$

Defined like this the second derivative is difficult to evaluate in practice as  $J''(u)w$  is an element of  $V'$ . Happily, by making it act on  $v \in V$  we obtain a continuous bilinear form over  $V \times V$  which we shall denote  $J''(u)(w, v)$  instead of  $(J''(u)w)v$ . We leave the reader the task of proving the following elementary result.

**Lemma 10.1.8** If  $J$  is a twice differentiable function from  $V$  into  $\mathbb{R}$ , it satisfies

$$J(u + w) = J(u) + J'(u)w + \frac{1}{2}J''(u)(w, w) + o(\|w\|^2), \quad \text{with } \lim_{w \rightarrow 0} \frac{o(\|w\|^2)}{\|w\|^2} = 0, \quad (10.14)$$

where  $J''(u)$  is identified with a continuous bilinear form over  $V \times V$ .

In practice this is the  $J''(u)(w, w)$  that we calculate.

**Exercise 10.1.8** Let  $a$  be a continuous symmetric bilinear form over  $V \times V$ . Let  $L$  be a continuous linear form over  $V$ . We set  $J(u) = \frac{1}{2}a(u, u) - L(u)$ . Show that  $J$  is twice differentiable over  $V$  and that  $J''(u)(v, w) = a(v, w)$  for all  $u, v, w \in V$ . Apply this result to the examples of exercises 10.1.3, 10.1.4, and 10.1.5.

When  $J$  is twice differentiable we recover the usual convexity condition: if the second derivative is positive, then the function is convex.

**Exercise 10.1.9** Show that if  $J$  is twice differentiable over  $V$  the conditions of propositions 10.1.4 and 10.1.5 are respectively equivalent to

$$J''(u)(w, w) \geq 0 \quad \text{and} \quad J''(u)(w, w) \geq \alpha \|w\|^2 \quad \forall u, w \in V. \quad (10.15)$$

## 10.2 Optimality conditions

### 10.2.1 Euler inequalities and convex constraints

We start by formulating the minimality conditions when the set of constraints  $K$  is convex, where things are more simple (we always assume that  $K$  is closed, nonempty, and that  $J$  is continuous over an open set containing  $K$ ). The essential idea of the result which follows is that, for all  $v \in K$ , we can test the optimality of  $u$  in the ‘admissible direction’  $(v - u)$  as  $u + h(v - u) \in K$  if  $h \in [0, 1]$ .

**Theorem 10.2.1 (Euler inequality, convex case)** *Let  $u \in K$  with  $K$  convex. We assume that  $J$  is differentiable at  $u$ . If  $u$  is a local minimum point of  $J$  over  $K$ , then*

$$\langle J'(u), v - u \rangle \geq 0 \quad \forall v \in K. \quad (10.16)$$

*If  $u \in K$  satisfies (10.16) and if  $J$  is convex, then  $u$  is a global minimum of  $J$  over  $K$ .*

**Remark 10.2.2** We call (10.16) a ‘Euler inequality’. This is a **necessary** condition for optimality which becomes **necessary and sufficient** if  $J$  is convex. We must also remark that, in two important cases, (10.16) **reduces simply to the Euler equation**  $J'(u) = 0$ . Initially, if  $K = V$ ,  $v - u$  produces the whole of  $V$  when  $v$  produces the whole of  $V$ , and therefore (10.16) implies  $J'(u) = 0$ . On the other hand, if  $u$  is in the interior of  $K$ , the same conclusion holds. •

**Proof.** For  $v \in K$  and  $h \in ]0, 1]$ ,  $u + h(v - u) \in K$ , and therefore

$$\frac{J(u + h(v - u)) - J(u)}{h} \geq 0. \quad (10.17)$$

We deduce (10.16) letting  $h$  tend to 0. The second assertion of the theorem follows immediately from (10.7).  $\square$

**Exercise 10.2.1** Let  $K$  be a convex closed nonempty set of  $V$ . For  $x \in V$ , we look for the projection  $x_K \in K$  of  $x$  over  $K$  (see theorem 12.1.10)

$$\|x - x_K\| = \min_{y \in K} \|x - y\|.$$

Show that the necessary and sufficient condition (10.16) reduces exactly to (12.1).



**Exercise 10.2.2** Let  $A$  be a real matrix of order  $p \times n$  and  $b \in \mathbb{R}^p$ . We consider the 'least-squares' problem

$$\inf_{x \in \mathbb{R}^n} \|Ax - b\|^2.$$

Show that this problem always has a solution and write the corresponding Euler equation.

**Exercise 10.2.3** We return to example 9.1.6

$$\inf_{x \in \text{Ker } B} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\}$$

with  $A$  a square symmetric matrix of order  $n$ , and  $B$  of size  $m \times n$  ( $m \leq n$ ). Show that there exists a unique solution if  $A$  is positive definite. Show that every minimum point  $\bar{x} \in \mathbb{R}^n$  satisfies

$$A\bar{x} - b = B^*p \quad \text{with } p \in \mathbb{R}^m.$$

**Exercise 10.2.4** We return to example 9.1.10. Show that the Euler equation satisfied by the minimum point  $u \in H_0^1(\Omega)$  of

$$\inf_{v \in H_0^1(\Omega)} \left\{ J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx \right\}$$

is precisely the variational formulation

$$\int_{\Omega} \nabla u \cdot \nabla v dx = \int_{\Omega} f v dx \quad \forall v \in H_0^1(\Omega).$$

(We therefore recover a result of proposition 5.2.7.)

**Exercise 10.2.5** Let  $K$  be a convex closed nonempty set of  $V$ , let  $a$  be a symmetric continuous bilinear coercive form over  $V$ , and let  $L$  be a continuous linear form over  $V$ . Show that  $J(v) = \frac{1}{2}a(v, v) - L(v)$  has a unique minimum point in  $K$ , denoted as  $u$ . Show that  $u$  is also the unique solution of the problem (called a variational inequality)

$$u \in K \quad \text{and} \quad a(u, v - u) \geq L(v - u) \quad \forall v \in K.$$

**Exercise 10.2.6** Let  $J_1$  and  $J_2$  be two continuous convex functions over a nonempty closed convex set  $K \subset V$ . We assume that  $J_1$  is only differentiable. Show that  $u \in K$  is a minimum of  $J_1 + J_2$  if and only if

$$\langle J_1'(u), v - u \rangle + J_2(v) - J_2(u) \geq 0 \quad \forall v \in K.$$

The following remarks, which are simple applications of theorem 10.2.1, will give us the intuition for the idea of a 'Lagrange multiplier' which will be developed in the following section.

**Remark 10.2.3** Let us examine the particular case where  $K$  is a closed affine subspace of  $V$ . We therefore assume that  $K = u_0 + \mathcal{P}$ , where  $u_0 \in V$  and where  $\mathcal{P}$  is a closed vector subspace of  $V$ . Then, when  $v$  describes  $K$ ,  $v - u$  is an arbitrary element of  $\mathcal{P}$  so that (10.16) is equivalent to

$$\langle J'(u), w \rangle = 0 \quad \forall w \in \mathcal{P},$$

that is to say  $J'(u) \in \mathcal{P}^\perp$ . In particular, if  $\mathcal{P}$  is a finite intersection of hyperplanes, that is to say if

$$\mathcal{P} = \{v \in V, \quad \langle a_i, v \rangle = 0 \quad \text{for } 1 \leq i \leq M\},$$

where  $a_1, \dots, a_m$  are given in  $V$ , then (the reader will verify that)  $\mathcal{P}^\perp$  is the vector space generated by the family  $(a_i)_{1 \leq i \leq M}$ . The optimality condition is therefore written in the form:

$$u \in K \quad \text{and} \quad \exists \lambda_1, \dots, \lambda_M \in \mathbb{R}, \quad J'(u) + \sum_{i=1}^M \lambda_i a_i = 0, \quad (10.18)$$

and the real numbers  $\lambda_i$  are called **Lagrange multipliers**. We see in theorem 10.2.8 their more general and fundamental role. •

**Remark 10.2.4** Let us now suppose that  $K$  is a closed convex cone, which means that  $K$  is a convex closed set such that  $\lambda v \in K$  for all  $v \in K$  and all  $\lambda \geq 0$ . By taking  $v = 0$  then  $v = 2u$  in (10.16), we obtain

$$\langle J'(u), u \rangle = 0. \quad (10.19)$$

Consequently, (10.16) implies that

$$\langle J'(u), w \rangle \geq 0 \quad \forall w \in K. \quad (10.20)$$

In fact, (10.16) is equivalent to (10.19) and (10.20). In the case where

$$K = \{v \in V, \quad \langle a_i, v \rangle \leq 0 \quad \text{for } 1 \leq i \leq M\},$$

where  $a_1, \dots, a_M$  are given in  $V$ , the Farkas lemma 10.2.17 (see below) shows that

$$u \in K \quad \text{and} \quad \exists \lambda_1, \dots, \lambda_M \geq 0, \quad J'(u) + \sum_{i=1}^M \lambda_i a_i = 0, \quad (10.21)$$

and the equality (10.19) shows that  $\lambda_i = 0$  if  $\langle a_i, u \rangle < 0$ . The nonnegative real numbers  $\lambda_i$  are again called **Lagrange multipliers**. We see in theorem 10.2.15 their more general and fundamental role. •

Let us finish this section by giving a **second order optimality condition**.

**Proposition 10.2.5** *We assume that  $K = V$  and that  $J$  is twice differentiable at  $u$ . If  $u$  is a local minimum point of  $J$ , then*

$$J'(u) = 0 \quad \text{and} \quad J''(u)(w, w) \geq 0 \quad \forall w \in V. \quad (10.22)$$

*Conversely, if, for all  $v$  in a neighbourhood of  $u$ ,*

$$J'(u) = 0 \quad \text{and} \quad J''(v)(w, w) \geq 0 \quad \forall w \in V, \quad (10.23)$$

*then  $u$  is a local minimum of  $J$ .*

**Proof.** If  $u$  is a local minimum point, we already know that  $J'(u) = 0$  and formula (10.14) gives us (10.22). Conversely, if  $u$  satisfies (10.23), we write a second order Taylor expansion (in the neighbourhood of zero) with exact remainder for the function  $\phi(t) = J(u + tw)$  with  $t \in \mathbb{R}$  and we deduce easily that  $u$  is a local minimum of  $J$  (see definition 9.1.1).  $\square$

## 10.2.2 Lagrange multipliers

We shall now try to write the minimality conditions when the set  $K$  is not convex. More precisely, we shall study sets  $K$  defined by **equality constraints** or **inequality constraints** (or both at the same time). We start with a general remark about **admissible directions**.

**Definition 10.2.6** *At every point  $v \in K$ , the set*

$$K(v) = \left\{ w \in V, \exists (v^n) \in K^{\mathbb{N}}, \exists (\varepsilon^n) \in (\mathbb{R}_+^*)^{\mathbb{N}}, \right. \\ \left. \lim_{n \rightarrow +\infty} v^n = v, \lim_{n \rightarrow +\infty} \varepsilon^n = 0, \lim_{n \rightarrow +\infty} (v^n - v)/\varepsilon^n = w \right\}$$

*is called the cone of admissible directions at the point  $v$ .*

In more visual terms, we can also say that  $K(v)$  is the set of all the vectors which are tangents at  $v$  with a curve contained in  $K$  and passing through  $v$  (if  $K$  is a regular variety,  $K(v)$  is simply the tangent space to  $K$  at  $v$ ). In other words,  $K(v)$  is the set of all the possible directions of variations starting from  $v$  which remain infinitesimally in  $K$ .

By setting  $w^n = (v^n - v)/\varepsilon^n$ , we can also say equivalently that  $w \in K(v)$  if and only if there exists a sequence  $w^n$  in  $V$  and a sequence  $\varepsilon^n$  in  $\mathbb{R}$  such that

$$\lim_{n \rightarrow +\infty} w^n = w, \quad \lim_{n \rightarrow +\infty} \varepsilon^n = 0, \quad \text{and} \quad v + \varepsilon^n w^n \in K \quad \forall n.$$

It is easy to verify that the set  $K(v)$  (which could well be reduced to  $\{0\}$ !) is a cone:  $\lambda w \in K(v)$  for all  $w \in K(v)$  and all  $\lambda \geq 0$ .

**Exercise 10.2.7** Show that  $K(v)$  is a closed cone and that  $K(v) = V$  if  $v$  is interior to  $K$ . Give an example where  $K(v)$  is reduced to  $\{0\}$ .

The interest in the cone of admissible directions lies in the following result, which gives a **necessary** optimality condition. The proof, which is very simple, is left to the reader.

**Proposition 10.2.7 (Euler inequality, general case)** *Let  $u$  be a local minimum of  $J$  over  $K$ . If  $J$  is differentiable at  $u$ , we have*

$$\langle J'(u), w \rangle \geq 0 \quad \forall w \in K(u).$$

We shall now make precise the necessary condition of proposition 10.2.7 in the case where  $K$  is given by **equality** or **inequality constraints**. The results that we will obtain generalize those of remarks 10.2.3 and 10.2.4.

### Equality constraints

In this first case we assume that  $K$  is given by

$$K = \{v \in V, \quad F(v) = 0\}, \quad (10.24)$$

where  $F(v) = (F_1(v), \dots, F_M(v))$  is a mapping from  $V$  into  $\mathbb{R}^M$ , with  $M \geq 1$ . The **necessary** optimality condition then takes the following form.

**Theorem 10.2.8** *Take  $u \in K$  where  $K$  is given by (10.24). We assume that  $J$  is differentiable at  $u \in K$  and that the functions  $F_i$  ( $1 \leq i \leq M$ ) are continuously differentiable in a neighbourhood of  $u$ . We further assume that the vectors  $(F'_i(u))_{1 \leq i \leq M}$  are linearly independent. Then, if  $u$  is a local minimum of  $J$  over  $K$ , there exist  $\lambda_1, \dots, \lambda_M \in \mathbb{R}$ , called **Lagrange multipliers**, such that*

$$J'(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0. \quad (10.25)$$

**Proof.** As the vectors  $(F'_i(u))_{1 \leq i \leq M}$  are linearly independent, the implicit function theorem allows us to show that

$$K(u) = \{w \in V, \quad \langle F'_i(u), w \rangle = 0 \quad \text{for } i = 1, \dots, M\}, \quad (10.26)$$

or equivalently

$$K(u) = \bigcap_{i=1}^M [F'_i(u)]^\perp. \quad (10.27)$$

We shall not detail the proof of (10.26) which is a classical result of differential calculus (in fact,  $K(u)$  is the tangent space to the variety  $K$  at the point  $u$ ). As  $K(u)$  is a vector space, we can successively take  $w$  and  $-w$  in proposition 10.2.7, which leads to

$$\langle J'(u), w \rangle = 0 \quad \forall w \in \bigcap_{i=1}^M [F'_i(u)]^\perp,$$

that is to say that  $J'(u)$  is generated by  $(F'_i(u))_{1 \leq i \leq M}$  (let us note that the Lagrange multipliers are defined uniquely). Another proof (which is more geometrical) is proposed in the proof of proposition 10.2.11.  $\square$

**Remark 10.2.9** When the vectors  $(F'_i(u))_{1 \leq i \leq M}$  are linearly independent, we say that we are in a **regular case**. In the opposite case, we talk of a **nonregular case** and the conclusion of theorem 10.2.8 is false as the following example shows.

Let us take  $V = \mathbb{R}$ ,  $M = 1$ ,  $F(v) = v^2$ ,  $J(v) = v$ , from which  $K = \{0\}$ ,  $u = 0$ ,  $F'(u) = 0$ : we therefore have a nonregular case. As  $J'(u) = 1$ , (10.25) does not hold.  $\bullet$

To fully understand the range of theorem 10.2.8, we apply it to example 9.1.6

$$\min_{x \in \text{Ker } B} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\},$$

where  $A$  is symmetric positive definite of order  $n$ , and  $B$  of size  $m \times n$  with  $m \leq n$ . We denote by  $(b_i)_{1 \leq i \leq m}$  the  $m$  rows of  $B$ . In this problem there are therefore  $m$  constraints  $b_i \cdot x = 0$ . If the range of  $B$  is  $m$ , the  $(b_i)$  are independent and we can apply the conclusion (10.25) (if not, the constraints are either redundant, or contradictory). There therefore exists a Lagrange multiplier  $p \in \mathbb{R}^m$  such that a minimum point  $\bar{x}$  satisfies

$$A\bar{x} - b = \sum_{i=1}^m p_i b_i = B^* p.$$

If  $A$  is invertible and if  $B$  has range  $m$ , we deduce the value of  $\bar{x}$ : we have  $\bar{x} = A^{-1}(b + B^* p)$ , and as  $B\bar{x} = 0$  we obtain

$$p = -(BA^{-1}B^*)^{-1} BA^{-1}b \quad \text{and} \quad \bar{x} = A^{-1} \left( I - B^* (BA^{-1}B^*)^{-1} BA^{-1} \right) b.$$

In particular, we have thus proved that the Lagrange multiplier  $p$  is unique if  $B$  has range  $m$ . If this is not the case, we know that there exists a solution of  $BA^{-1}B^* p = -BA^{-1}b$ ,  $p$ , which is only unique up to the addition of a vector in the kernel of  $B^*$ . We therefore recover the result of exercise 10.2.3.

**Exercise 10.2.8** Generalize the results above for this variant of example 9.1.6

$$\min_{Bx=c} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\},$$

where  $c \in \mathbb{R}^m$  is a given vector.

**Exercise 10.2.9** Apply theorem 10.2.8 to example 9.1.7 and deduce that the minimum points of  $J$  over the unit sphere are eigenvectors of  $A$  associated with the smallest eigenvalue.

**Exercise 10.2.10** By using the preceding results and those of exercise 10.1.6, show that the solution of Didon's problem (example 9.1.11) is necessarily an arc of a circle.

**Exercise 10.2.11** We study the first eigenvalue of the Laplacian in a bounded domain  $\Omega$  (see Section 7.3). For this we introduce the minimization problem over  $K = \{v \in H_0^1(\Omega), \int_{\Omega} v^2 dx = 1\}$

$$\min_{v \in K} \left\{ J(v) = \int_{\Omega} |\nabla v|^2 dx \right\}.$$

Show that this problem has a minimum (we shall show that  $K$  is compact for the minimizing sequences with the help of Rellich's theorem 4.3.21). Write the Euler equation for this problem and deduce that the value of the minimum is the first eigenvalue and that the minimum points are the associated eigenvectors.

**Exercise 10.2.12** Let  $A$  be an  $n \times n$  symmetric positive definite matrix and  $b \in \mathbb{R}^n$  nonzero.

1. Show that the problems

$$\sup_{Ax \cdot x \leq 1} b \cdot x \quad \text{and} \quad \sup_{Ax \cdot x = 1} b \cdot x$$

are equivalent and that they have a solution. Use the theorem 10.2.8 to calculate this solution and show that it is unique.

2. We introduce a partial order in the set of symmetric positive definite matrices of order  $n$  by saying that  $A \geq B$  if and only if  $Ax \cdot x \geq Bx \cdot x$  for all  $x \in \mathbb{R}^n$ . Deduce from the question above that, if  $A \geq B$ , then  $B^{-1} \geq A^{-1}$ .

**Exercise 10.2.13** In the kinetic theory of gas, the molecules of gas are represented at every point of the space by a distribution function  $f(v)$  depending on the microscopic velocity  $v \in \mathbb{R}^N$ . The macroscopic quantities, like the density of the gas  $\rho$ , its velocity  $u$ , and its temperature  $T$ , are recovered thanks to the moments of the function  $f(v)$

$$\rho = \int_{\mathbb{R}^N} f(v) dv, \quad \rho u = \int_{\mathbb{R}^N} v f(v) dv, \quad \frac{1}{2} \rho u^2 + \frac{N}{2} \rho T = \frac{1}{2} \int_{\mathbb{R}^N} |v|^2 f(v) dv. \quad (10.28)$$

Boltzmann introduced the kinetic entropy  $H(f)$  defined by

$$H(f) = \int_{\mathbb{R}^N} f(v) \log(f(v)) dv.$$

Show that  $H$  is strictly convex over the space of measurable functions  $f(v) > 0$  such that  $H(f) < +\infty$ . We minimize  $H$  over this space under the moment constraints moment (10.28), and we will find that there exists a unique minimum point  $M(v)$ . Show that this minimum point is a Maxwellian defined by

$$M(v) = \frac{\rho}{(2\pi T)^{N/2}} \exp\left(-\frac{|v - u|^2}{2T}\right).$$

**Remark 10.2.10** It is useful to introduce the function  $\mathcal{L}$  defined over  $V \times \mathbb{R}^M$  by

$$\mathcal{L}(v, \mu) = J(v) + \sum_{i=1}^M \mu_i F_i(v) = J(v) + \mu \cdot F(v)$$

that we call the **Lagrangian** of the minimization problem of  $J$  over  $K$ . If  $u \in K$  is a local minimum of  $J$  over  $K$ , theorem 10.2.8 then tells us that, in the regular case, there exists a  $\lambda \in \mathbb{R}^M$  such that

$$\frac{\partial \mathcal{L}}{\partial v}(u, \lambda) = 0, \quad \frac{\partial \mathcal{L}}{\partial \mu}(u, \lambda) = 0,$$

since  $\frac{\partial \mathcal{L}}{\partial \mu}(u, \lambda) = F(u) = 0$  if  $u \in K$  and  $\frac{\partial \mathcal{L}}{\partial v}(u, \lambda) = J'(u) + \lambda F'(u) = 0$  from (10.25). We can therefore write the constraint and the optimality condition as the annihilation of the gradient (the stationarity) of the Lagrangian. •

We now give a **necessary** second order optimality condition.

**Proposition 10.2.11** *We take the hypotheses of theorem 10.2.8 and we assume that the functions  $J$  and  $F_1, \dots, F_M$  are twice continuously differentiable and that the vectors  $(F'_i(u))_{1 \leq i \leq M}$  are linearly independent. Let  $\lambda \in \mathbb{R}^M$  be the Lagrange multiplier defined by theorem 10.2.8. Then every local minimum  $u$  of  $J$  over  $K$  satisfies*

$$\left( J''(u) + \sum_{i=1}^M \lambda_i F''_i(u) \right) (w, w) \geq 0 \quad \forall w \in K(u) = \bigcap_{i=1}^M [F'_i(u)]^\perp. \quad (10.29)$$

**Proof.** Let us suppose that there exists an admissible function of class  $C^2$ , that is to say a function  $t \rightarrow u(t)$  of  $[0, 1]$  in  $V$  such that  $u(0) = u$  and  $F(u(t)) = 0$  for all  $t \in [0, 1]$ . By definition, the derivative  $u'(0)$  belongs to the cone of admissible directions  $K(u)$ . We set

$$j(t) = J(u(t)) \quad \text{and} \quad f_i(t) = F_i(u(t)) \quad \text{for } 1 \leq i \leq M.$$

By differentiating we obtain

$$j'(t) = \langle J'(u(t)), u'(t) \rangle \quad \text{and} \quad f'_i(t) = \langle F'_i(u(t)), u'(t) \rangle \quad \text{for } 1 \leq i \leq M,$$

and

$$\begin{aligned} j''(t) &= J''(u(t))(u'(t), u'(t)) + \langle J'(u(t)), u''(t) \rangle \\ f''_i(t) &= F''_i(u(t))(u'(t), u'(t)) + \langle F'_i(u(t)), u''(t) \rangle \quad \text{for } 1 \leq i \leq M. \end{aligned}$$

As  $f_i(t) = 0$  for all  $t$  and since 0 is a minimum of  $j(t)$ , we deduce  $j'(0) = 0$ ,  $j''(0) \geq 0$ , and  $f'_i(0) = f''_i(0) = 0$ . The conditions  $f'_i(0) = 0$  tell us that  $u'(0)$  is orthogonal to

the subspace generated by  $(F'_i(u))_{1 \leq i \leq M}$  (which is equal to  $K(u)$  when this family is linearly independent), while  $j'(0) = 0$  means that  $J'(u)$  is orthogonal to  $u'(0)$ . If  $u'(0)$  describes all of  $K(u)$  when we vary the admissible functions, we deduce that  $J'(u)$  and the  $F'_i(u)$  belongs to the same subspace (the orthogonal space of  $K(u)$ ). We therefore recover the first order condition: there exists  $\lambda \in \mathbb{R}^M$  such that

$$J'(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0. \quad (10.30)$$

Conditions  $f''_i(0) = 0$  imply that

$$0 = \sum_{i=1}^M \lambda_i \left( F''_i(u)(u'(0), u'(0)) + \langle F'_i(u), u''(0) \rangle \right),$$

while  $j''(0) \geq 0$  gives

$$J''(u)(u'(0), u'(0)) + \langle J'(u), u''(0) \rangle \geq 0.$$

Thanks to (10.30) we can eliminate the first derivatives and  $u''(0)$  to obtain (by summing the two last equations)

$$\left( \sum_{i=1}^M \lambda_i F''_i(u) + J''(u) \right) (u'(0), u'(0)) \geq 0,$$

which is none other than (10.29) when  $u'(0)$  generates  $K(u)$ .

The existence of such admissible functions  $u(t)$  and the fact that the set of  $u'(0)$  describes the entire cone of admissible directions  $K(u)$  is a consequence of the implicit function theorem that we can apply thanks to the hypothesis that the family  $(F'_i(u))_{1 \leq i \leq M}$  is linearly independent (we leave the technical details to the courageous reader see Chapter 9 in [19]).  $\square$

**Exercise 10.2.14** Calculate the second order necessary optimality condition for the examples 9.1.6 and 9.1.7.

### Inequality constraints

In this second case we assume that  $K$  is given by

$$K = \{v \in V, \quad F_i(v) \leq 0 \quad \text{for } 1 \leq i \leq M\}, \quad (10.31)$$

where  $F_1, \dots, F_M$  are always functions from  $V$  into  $\mathbb{R}$ . When we want to determine the cone of admissible directions  $K(v)$ , the situation is a little more complicated than before as all the constraints in (10.31) do not play the same role depending on the point  $v$  where we calculate  $K(v)$ . In effect, if  $F_i(v) < 0$ , it is clear that, for  $\epsilon$  sufficiently



small, we will also have  $F_i(v + \epsilon w) \leq 0$  (we say that the constraint  $i$  is inactive at  $v$ ). If  $F_i(v) = 0$  for certain indices  $i$ , it is not clear that we can find a vector  $w \in V$  such that, for  $\epsilon > 0$  sufficiently small,  $(v + \epsilon w)$  satisfies all the constraints in (10.31). It will therefore be necessary to impose supplementary conditions on the constraints, called **constraint qualifications**. Roughly speaking, these conditions will guarantee that we can make ‘variations’ around a point  $v$  in order to test its optimality. There exist different types of constraint qualification (more or less sophisticated and general). We shall give a definition whose principle is to look at the **linearized** problem if it is possible to make variations respecting the linearized constraints. These ‘calculus of variations’ considerations motivate the following definitions.

**Definition 10.2.12** Take  $u \in K$ . The set  $I(u) = \{i \in \{1, \dots, M\}, F_i(u) = 0\}$  is called the set of **active** constraints at  $u$ .

**Definition 10.2.13** We say that the constraints (10.31) are **qualified** at  $u \in K$  if and only if there exists a direction  $\bar{w} \in V$  such that we have for all  $i \in I(u)$

$$\begin{aligned} &\text{either } \langle F'_i(u), \bar{w} \rangle < 0, \\ &\text{or } \langle F'_i(u), \bar{w} \rangle = 0, \quad \text{and } F_i \text{ is affine.} \end{aligned} \quad (10.32)$$

**Remark 10.2.14** The direction  $\bar{w}$  is in some way a ‘re-entrant direction’ since we deduce from (10.32) that  $u + \epsilon \bar{w} \in K$  for all  $\epsilon \geq 0$  sufficiently small. Of course, if all the functions  $F_i$  are affine, we can take  $\bar{w} = 0$  and the constraints are automatically qualified. The reasoning for distinguishing the affine constraints in Definition 10.2.13 is justified not only because those are qualified under less strict conditions, but above all because of the importance of affine constraints in applications (as we see in the examples of Chapter 9). •

We can then state the **necessary** optimality conditions over the set (10.31).

**Theorem 10.2.15** We assume that  $K$  is given by (10.31), that the functions  $J$  and  $F_1, \dots, F_M$  are differentiable at  $u$  and that the constraints are qualified at  $u$ . Then, if  $u$  is a local minimum of  $J$  over  $K$ , there exist  $\lambda_1, \dots, \lambda_M \geq 0$ , called *Lagrange multipliers*, such that

$$J'(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0, \quad \lambda_i \geq 0, \quad \lambda_i = 0 \text{ if } F_i(u) < 0 \quad \forall i \in \{1, \dots, M\}. \quad (10.33)$$

**Remark 10.2.16** We can rewrite the condition (10.33) in the following form

$$J'(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0, \quad \lambda \geq 0, \quad \lambda \cdot F(u) = 0,$$

where  $\lambda \geq 0$  means that each of the components of the vector  $\lambda = (\lambda_1, \dots, \lambda_M)$  is positive, since, for every index  $i \in \{1, \dots, M\}$ , we have either  $F_i(u) = 0$ , or  $\lambda_i = 0$ . The fact that  $\lambda \cdot F(u) = 0$  is called the condition of complementary variations. •

**Proof.** Let us consider first of all the set

$$\tilde{K}(u) = \{w \in V, \quad \langle F'_i(u), w \rangle \leq 0 \quad \forall i \in I(u)\}. \quad (10.34)$$

(We can show that  $\tilde{K}(u)$  is none other than the cone  $K(u)$  of admissible directions). Let  $\bar{w}$  be an admissible direction satisfying (10.32),  $w \in \tilde{K}(u)$ , and take a real number  $\delta > 0$ . We shall show that  $u + \varepsilon(w + \delta\bar{w}) \in K$  for every real number  $\varepsilon > 0$  which is small enough. We must examine three important cases.

1. If  $i \notin I(u)$ , we have  $F_i(u) < 0$  and  $F_i(u + \varepsilon(w + \delta\bar{w})) < 0$  by continuity if  $\varepsilon$  is small enough.
2. If  $i \in I(u)$  and  $\langle F'_i(u), \bar{w} \rangle < 0$ , then

$$\begin{aligned} F_i(u + \varepsilon(w + \delta\bar{w})) &= F_i(u) + \varepsilon \langle F'_i(u), w + \delta\bar{w} \rangle + o(\varepsilon) \\ &\leq \varepsilon \delta \langle F'_i(u), \bar{w} \rangle + o(\varepsilon), \end{aligned} \quad (10.35)$$

from which we have  $F_i(u + \varepsilon(w + \delta\bar{w})) < 0$  for  $\varepsilon > 0$  small enough.

3. Finally, if  $i \in I(u)$  and  $\langle F'_i(u), \bar{w} \rangle = 0$ , then  $F_i$  is affine and

$$F_i(u + \varepsilon(w + \delta\bar{w})) = F_i(u) + \varepsilon \langle F'_i(u), w + \delta\bar{w} \rangle = \varepsilon \langle F'_i(u), w \rangle \leq 0. \quad (10.36)$$

Finally, if  $u$  is a local minimum of  $J$  over  $K$ , we deduce from above that

$$\langle J'(u), w + \delta\bar{w} \rangle \geq 0 \quad \forall w \in \tilde{K}(u), \quad \forall \delta \in \mathbb{R}_+^*.$$

This implies that  $\langle J'(u), w \rangle \geq 0 \quad \forall w \in \tilde{K}(u)$  and we finish the proof thanks to the Farkas lemma 10.2.17 below.  $\square$

**Lemma 10.2.17 (Farkas)** *Let  $a_1, \dots, a_M$  be fixed elements of  $V$ . We consider the sets*

$$\mathcal{K} = \left\{ v \in V, \quad \langle a_i, v \rangle \leq 0 \quad \text{for } 1 \leq i \leq M \right\},$$

and

$$\hat{\mathcal{K}} = \left\{ v \in V, \quad \exists \lambda_1, \dots, \lambda_M \geq 0, \quad v = - \sum_{i=1}^M \lambda_i a_i \right\}.$$

Then for every  $p \in V$ , we have the implication

$$\langle p, w \rangle \geq 0 \quad \forall w \in \mathcal{K} \implies p \in \hat{\mathcal{K}}.$$

(The reciprocal being obvious, it is in fact an equivalence.)

**Proof.** Let us start by showing that  $\hat{\mathcal{K}}$  is closed. Since this property is obvious when  $M = 1$ , we proceed by recurrence and assume that it is true when the number of vectors  $a_i$  is less than  $M$ .

Let us suppose first that the vectors  $(a_i)_{1 \leq i \leq M}$  are linearly independent. Let  $(v^n) = \left(-\sum_{i=1}^M \lambda_i^n a_i\right)$  be a sequence of elements of  $\hat{\mathcal{K}}$  (therefore with  $\lambda_i^n \geq 0 \forall i \forall n$ ), convergent towards a limit  $v \in V$ . Then it is clear that each sequence  $(\lambda_i^n)$  converges in  $\mathbb{R}_+$  to a limit  $\lambda_i \geq 0$  (for  $1 \leq i \leq M$ ) since the vectors  $(a_i)_{1 \leq i \leq M}$  form a basis of the space that they generate. We therefore have  $v = -\sum_{i=1}^M \lambda_i a_i \in \hat{\mathcal{K}}$ , which is therefore closed.

If the vectors  $(a_i)_{1 \leq i \leq M}$  are linearly dependent, there exists a relation of the form  $\sum_{i=1}^M \mu_i a_i = 0$ , and we can assume that at least one of the coefficients  $\mu_i$  is strictly positive. Then let  $v = -\sum_{i=1}^M \lambda_i a_i$  be an element of  $\hat{\mathcal{K}}$ . For all  $t \leq 0$ , we can also write  $v = -\sum_{i=1}^M (\lambda_i + t\mu_i) a_i$ , and we can choose  $t \leq 0$  so that

$$\lambda_i + t\mu_i \geq 0 \quad \forall i \in \{1, \dots, M\} \quad \text{and} \quad \exists i_0 \in \{1, \dots, M\}, \quad \lambda_{i_0} + t\mu_{i_0} = 0.$$

This reasoning shows that

$$\hat{\mathcal{K}} = \bigcup_{i_0=1}^M \left\{ v \in V, \exists \lambda_1, \dots, \lambda_M \geq 0, \quad v = -\sum_{i \neq i_0} \lambda_i a_i \right\}. \quad (10.37)$$

By our recurrence hypothesis, each of the sets appearing in the right-hand side of (10.37) is closed, and it is therefore the same for  $\hat{\mathcal{K}}$ .

Let us now reason by contradiction: let us assume that  $\langle p, w \rangle \geq 0 \forall w \in \mathcal{K}$  and that  $p \notin \hat{\mathcal{K}}$ . We can then use theorem 12.1.19 of the separation of a point and a convex set to separate  $p$  and  $\hat{\mathcal{K}}$  which is closed and, obviously, convex and nonempty. Therefore there exist  $w \neq 0$  in  $V$  and  $\alpha \in \mathbb{R}$  such that

$$\langle p, w \rangle < \alpha < \langle w, v \rangle \quad \forall v \in \hat{\mathcal{K}}. \quad (10.38)$$

But then, we must have  $\alpha < 0$  since  $0 \in \hat{\mathcal{K}}$ ; on the other hand, for all  $i \in \{1, \dots, M\}$  we can choose, in (10.38),  $v = -\lambda a_i$  with  $\lambda$  arbitrarily large, which shows that  $\langle w, a_i \rangle \leq 0$ . We therefore obtain that  $w \in \mathcal{K}$  and that  $\langle p, w \rangle < \alpha < 0$ , which is impossible.  $\square$

**Exercise 10.2.15** Let  $A$  be a symmetric positive definite matrix of order  $n$ , and  $B$  a matrix of size  $m \times n$  with  $m \leq n$  and of range  $m$ . We consider the minimization problem

$$\min_{x \in \mathbb{R}^n, Bx \leq c} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\},$$

Apply theorem 10.2.15 to obtain the existence of a Lagrange multiplier  $p \in \mathbb{R}^m$  such that a minimum point  $\bar{x}$  satisfies

$$A\bar{x} - b + B^*p = 0, \quad p \geq 0, \quad p \cdot (B\bar{x} - c) = 0.$$

**Exercise 10.2.16** Let  $f \in L^2(\Omega)$  be a function defined over a bounded open set  $\Omega$ . For  $\epsilon > 0$  we consider the following regularization problem

$$\min_{u \in H_0^1(\Omega), \|u-f\|_{L^2(\Omega)} \leq \epsilon} \int_{\Omega} |\nabla u|^2 dx.$$

Show that this problem has a unique solution  $u_{\epsilon}$ . Show that, whether  $u_{\epsilon} = f$ , or  $u_{\epsilon} = 0$ , or there exists  $\lambda > 0$  such that  $u_{\epsilon}$  is the solution of

$$\begin{cases} -\Delta u_{\epsilon} + \lambda(u_{\epsilon} - f) = 0 & \text{in } \Omega, \\ u_{\epsilon} = 0 & \text{on } \partial\Omega. \end{cases}$$

### Equality and inequality constraints

We can of course mix the two types of constraints. We therefore assume that  $K$  is given by

$$K = \{v \in V, \quad G(v) = 0, \quad F(v) \leq 0\}, \quad (10.39)$$

where  $G(v) = (G_1(v), \dots, G_N(v))$  and  $F(v) = (F_1(v), \dots, F_M(v))$  are two mappings from  $V$  into  $\mathbb{R}^N$  and  $\mathbb{R}^M$ . In this new context, we must give an adequate definition of the qualification of constraints. We always denote by  $I(u) = \{i \in \{1, \dots, M\}, F_i(u) = 0\}$  the set of active inequality constraints at  $u \in K$ .

**Definition 10.2.18** We say that the constraints (10.39) are **qualified** at  $u \in K$  if and only if the vectors  $(G'_i(u))_{1 \leq i \leq N}$  are linearly independent and there exists a direction  $\bar{w} \in \bigcap_{i=1}^N [G'_i(u)]^{\perp}$  such that we have for all  $i \in I(u)$

$$\langle F'_i(u), \bar{w} \rangle < 0. \quad (10.40)$$

We can then state the **necessary** optimality conditions over the set (10.39).

**Theorem 10.2.19** Take  $u \in K$  where  $K$  is given by (10.39). We assume that  $J$  and  $F$  are differentiable at  $u$ , that  $G$  is differentiable in a neighbourhood of  $u$ , and that the constraints are qualified at  $u$  (in the sense of definition 10.2.18). Then, if  $u$  is a local minimum of  $J$  over  $K$ , there exist Lagrange multipliers  $\mu_1, \dots, \mu_N$ , and  $\lambda_1, \dots, \lambda_M \geq 0$ , such that

$$J'(u) + \sum_{i=1}^N \mu_i G'_i(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0, \quad \lambda \geq 0, \quad F(u) \leq 0, \quad \lambda \cdot F(u) = 0. \quad (10.41)$$

The proof of theorem 10.2.19 is a simple adaptation of that of theorems 10.2.8 and 10.2.15, which we leave to the reader as an exercise.

### Other forms of qualification conditions

Qualification conditions are **sufficient** conditions which are ‘geometrical’: they allow us to make internal variations in the set  $K$  starting from a point  $u \in K$ . The qualification condition of definition 10.2.13 is general enough (though far from being necessary), but sometimes difficult to verify in applications. This is why the remarks which follow give simpler (therefore easier to verify in practice) but less general qualification conditions (that is, less often satisfied).

**Remark 10.2.20** In the case of inequality constraints, we can take as a starting point the regular case (introduced in remark 10.2.9 for the equality constraints) in order to give a very simple condition which implies the qualification condition of definition 10.2.13. In effect, for  $u \in K$  the inactive constraints do not ‘play’ and only the active constraints  $i \in I(u)$  are taken into account which are just the equality constraints at this point! We can then easily check the following condition (which says that  $u$  is a regular point for the equality constraints  $F_i(u) = 0$  for  $i \in I(u)$ )

$$(F'_i(u))_{i \in I(u)} \text{ is a linearly independent family,} \quad (10.42)$$

which implies (10.32), that is to say that the constraints are qualified. In effect, it is enough to take  $\bar{w} = \sum_{i \in I(u)} \alpha_i F'_i(u)$  such that  $\langle F'_j(u), \bar{w} \rangle = -1$  for all  $j \in I(u)$  (the existence of coefficients  $\alpha_i$  follows from the invertibility of the matrix  $(\langle F'_i(u), F'_j(u) \rangle)_{ij}$ ). It is clear however that (10.32) does not imply (10.42). •

**Remark 10.2.21** In the case of combined equality and inequality constraints, we can also start from the regular case to give a simpler condition which implies the qualification condition of definition 10.2.18. This ‘strong’ (that is to say less often satisfied) qualification condition is

$$(G'_i(u))_{1 \leq i \leq N} \bigcup (F'_i(u))_{i \in I(u)} \text{ is a linearly independent family.} \quad (10.43)$$

We can easily check that (10.43) implies (10.40), that is to say the constraints are qualified. •

**Remark 10.2.22** Returning to the case of inequality constraints, which we assume to be convex, another possible qualification condition is the following. We assume that there exists  $\bar{v} \in V$  such that we have, for all  $i \in \{1, \dots, M\}$ ,

$$\begin{aligned} &\text{the functions } F_i \text{ are convex and,} \\ &\text{either } F_i(\bar{v}) < 0, \\ &\text{or } F_i(\bar{v}) = 0 \text{ and } F_i \text{ is affine.} \end{aligned} \quad (10.44)$$

The hypothesis (10.44) implies that the constraints are qualified on  $u \in K$  in the sense of definition 10.2.13. In effect, if  $i \in I(u)$  and if  $F_i(\bar{v}) < 0$ , then, from the condition of convexity (10.7)

$$\langle F'_i(u), \bar{v} - u \rangle = F_i(u) + \langle F'_i(u), \bar{v} - u \rangle \leq F_i(\bar{v}) < 0.$$

On the other hand, if  $i \in I(u)$  and if  $F_i(\bar{v}) = 0$ , then  $F_i$  is affine and

$$\langle F'_i(u), \bar{v} - u \rangle = F_i(\bar{v}) - F_i(u) = 0,$$

and definition 10.2.13 of qualification of constraints is satisfied with  $\bar{w} = \bar{v} - u$ . The advantage of hypothesis (10.44) is that we neither need to know the minimum point  $u$  nor calculate the derivatives of the functions  $F_1, \dots, F_M$ . •

## 10.3 Saddle point, Kuhn–Tucker theorem, duality

We have seen in remark 10.2.10 how it is possible to interpret the couple  $(u, \lambda)$  (minimum point, Lagrange multiplier) as a **stationary point of the Lagrangian**  $\mathcal{L}$ . We shall, in this section, make precise the nature of this stationary point as a **saddle point** and show how this formulation allows us to characterize a minimum (we shall see that, under certain hypotheses, the **necessary** conditions of stationarity of the Lagrangian are also **sufficient**). We shall briefly explore the **duality theory** which follows from this.

In addition to the theoretical interest of this characterization, its practical interest from the point of view of numerical algorithms will be illustrated in Section 10.5. Let us finally point out that the concept of saddle point plays a fundamental role in **game theory**.

### 10.3.1 Saddle point

Abstractly,  $V$  and  $Q$  are two real Hilbert spaces, a Lagrangian  $\mathcal{L}$  is a mapping of  $V \times Q$  (or of a part  $U \times P$  of  $V \times Q$ ) into  $\mathbb{R}$ . In the framework of theorem 10.2.8 (equality constraints), we had  $U = K$ ,  $P = Q = \mathbb{R}^M$  and  $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$ . The situation is a little different in the framework of theorem 10.2.15 (inequality constraints), where we had  $U = K$ ,  $Q = \mathbb{R}^M$ ,  $P = (\mathbb{R}_+)^M$  and again  $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$ .

Let us now give the definition of a saddle point.

**Definition 10.3.1** *We say that  $(u, p) \in U \times P$  is a saddle point of  $\mathcal{L}$  over  $U \times P$  if*

$$\forall q \in P \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall v \in U. \quad (10.45)$$

The following result shows the link between this notion of saddle point and the minimization problems with equality constraints (10.24) or inequality constraints (10.31) studied in the preceding section. For simplicity, we shall again use inequality between vectors, sometimes writing  $q \geq 0$  instead of  $q \in (\mathbb{R}_+)^M$ .

**Proposition 10.3.2** *We assume that the functions  $J, F_1, \dots, F_M$  are continuous over  $V$ , and that the set  $K$  is defined by (10.24) or (10.31). We denote by  $P = \mathbb{R}^M$  in the case of equality constraints (10.24) and  $P = (\mathbb{R}_+)^M$  in the case of inequality constraints (10.31). Let  $U$  be an open set of  $V$  containing  $K$ . For  $(v, q) \in U \times P$ , we set  $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$ .*

Let  $(u, p)$  be a saddle point of  $\mathcal{L}$  over  $U \times P$ . Then  $u \in K$  and  $u$  is a global minimum of  $J$  over  $K$ . Further, if  $J$  and  $F_1, \dots, F_M$  are differentiable at  $u$ , we have

$$J'(u) + \sum_{i=1}^M p_i F'_i(u) = 0. \quad (10.46)$$

**Proof.** Let us write the saddle point condition

$$\forall q \in P \quad J(u) + q \cdot F(u) \leq J(u) + p \cdot F(u) \leq J(v) + p \cdot F(v) \quad \forall v \in U. \quad (10.47)$$

Let us examine first the case of equality constraints. Since  $P = \mathbb{R}^M$ , the first inequality in (10.47) shows that  $F(u) = 0$ , that is,  $u \in K$ . Then  $J(u) \leq J(v) + p \cdot F(v) \quad \forall v \in U$ , which shows (by taking  $v \in K$ ) that  $u$  is a global minimum of  $J$  over  $K$ .

In the case of inequality constraints, we have  $P = (\mathbb{R}_+)^M$  and the first inequality of (10.47) shows now that  $F(u) \leq 0$  and that  $p \cdot F(u) = 0$ . This proves again that  $u \in K$ , and allows us to deduce easily from the second inequality that  $u$  is a global minimum of  $J$  over  $K$ .

Finally, if  $J$  and  $F_1, \dots, F_M$  are differentiable at  $u$ , the second inequality of (10.47) shows that  $u$  is a minimum point without constraint of  $J + p \cdot F$  in the open set  $U$ , which implies that the derivative is zero at  $u$ ,  $J'(u) + p \cdot F'(u) = 0$  (cf. remark 10.2.2).  $\square$

### 10.3.2 The Kuhn–Tucker theorem

We return to the minimization problem under inequality constraints for which the set  $K$  is given by (10.31), that is to say

$$K = \{v \in V, \quad F_i(v) \leq 0 \quad \text{for } 1 \leq i \leq m\}. \quad (10.48)$$

Theorem 10.2.15 has given a necessary condition for optimality. In this section we shall see that this condition is also **sufficient** if the constraints and the cost function are **convex**. In effect, proposition 10.3.2 says that, if  $(u, p)$  is a saddle point of the Lagrangian, then  $u$  realizes the minimum of  $J$  over  $K$ . For a convex minimization problem with convex inequality constraints, we shall establish a reciprocal of this result, that is to say that, if  $u$  realizes the minimum of  $J$  over  $K$ , then there exists  $p \in (\mathbb{R}_+)^M$  such that  $(u, p)$  is a saddle point of the Lagrangian. We assume from now on that  $J, F_1, \dots, F_M$  are convex and continuous over  $V$ .

**Remark 10.3.3** As  $J, F_1, \dots, F_M$  are convex and continuous,  $K$  is convex and closed and the existence of a global minimum of  $J$  over  $K$  is assured by theorem 9.2.7 when  $K$  is nonempty and that the condition (9.3) (we say ‘infinite at infinity’) is satisfied.  $\bullet$

The Kuhn–Tucker theorem (also sometimes called the theorem of Karush, Kuhn, and Tucker) confirms that, in the convex case, the necessary optimality condition of theorem 10.2.15 is in fact a **necessary and sufficient** condition.

**Theorem 10.3.4 (Kuhn–Tucker)** *We assume that the functions  $J, F_1, \dots, F_M$  are convex, continuous over  $V$  and differentiable over the set  $K$  (10.48). We introduce the associated Lagrangian  $\mathcal{L}$*

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v) \quad \forall (v, q) \in V \times (\mathbb{R}_+)^M.$$

*Let  $u \in K$  be a point of  $K$  where the constraints are qualified in the sense of definition 10.2.13. Then  $u$  is a global minimum of  $J$  over  $K$  if and only if there exists  $p \in (\mathbb{R}_+)^M$  such that  $(u, p)$  is a saddle point of the Lagrangian  $\mathcal{L}$  over  $V \times (\mathbb{R}_+)^M$  or, equivalently, such that*

$$F(u) \leq 0, \quad p \geq 0, \quad p \cdot F(u) = 0, \quad J'(u) + \sum_{i=1}^M p_i F'_i(u) = 0. \quad (10.49)$$

**Proof.** If  $u$  is a minimum of  $J$  over  $K$ , we can apply theorem 10.2.15, which exactly gives the optimality condition (10.49), from which we easily deduce that  $(u, p)$  is a saddle point of  $\mathcal{L}$  over  $V \times (\mathbb{R}_+)^M$  (by using the fact that  $J(v) + p \cdot F(v)$  is convex). Conversely, if  $(u, p)$  is a saddle point, we have already shown in proposition 10.3.2 that  $u$  is a global minimum of  $J$  over  $K$ .  $\square$

**Remark 10.3.5** The Kuhn–Tucker theorem 10.3.4 is only applied to inequality constraints, and not to equality constraints, in general. However, it is good to remark that **affine equality** constraints  $Av = b$  can be written in the form of inequality constraints (affine therefore convex)  $Av - b \leq 0$  and  $b - Av \leq 0$ . This allows us to apply the Kuhn–Tucker theorem 10.3.4 to the minimization problem with affine equality constraints.  $\bullet$

The following exercise allows us to interpret the Lagrange multipliers  $p_i$  as the sensitivity of the minimal value of  $J$  to variations of the constraints  $F_i$ : in economics, these coefficients measure prices or marginal costs, in mechanics forces corresponding to kinematic constraints, etc.

**Exercise 10.3.1** We consider the perturbed optimization problem

$$\inf_{F_i(v) \leq u_i, 1 \leq i \leq m} J(v), \quad (10.50)$$

with  $u_1, \dots, u_m \in \mathbb{R}$ . We use the hypotheses of the Kuhn–Tucker theorem 10.3.4. We denote by  $m^*(u)$  the minimal value of the perturbed problem (10.50).

1. Show that if  $p$  is the Lagrange multiplier for the nonperturbed problem (that is to say (10.50) with  $u = 0$ ), then

$$m^*(u) \geq m^*(0) - pu. \quad (10.51)$$

2. Deduce from (10.51) that if  $u \mapsto m^*(u)$  is differentiable, then

$$p_i = -\frac{\partial m^*}{\partial u_i}(0).$$

Interpret this result (cf. example 9.1.8 in economics).



### 10.3.3 Duality

Let us give a brief outline of the duality theory for the optimization problems. We shall apply it to the convex minimization problem with the inequality constraints in the preceding subsection. We have associated with this minimization problem a problem of finding a saddle point  $(u, p)$  for the Lagrangian  $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$ . But we shall see that, to the existence of a saddle point  $(u, p)$  of the Lagrangian, we can inversely associate not one but **two** optimization problems (more precisely, a minimization problem and a maximization problem), which will be called **duals** of one another. We will then explain using two simple examples how the introduction of the **dual problem** can be useful for the solution of the original problem, called the **primal problem** (as opposed to the dual).

Let us return for the moment to the general framework of definition 10.3.1.

**Definition 10.3.6** *Let  $V$  and  $Q$  be two real Hilbert spaces, and  $\mathcal{L}$  a Lagrangian defined over a subset  $U \times P$  of  $V \times Q$ . We assume that there exists a saddle point  $(u, p)$  of  $\mathcal{L}$  over  $U \times P$*

$$\forall q \in P \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall v \in U. \quad (10.52)$$

For  $v \in U$  and  $q \in P$ , let us set

$$\mathcal{J}(v) = \sup_{q \in P} \mathcal{L}(v, q) \quad \mathcal{G}(q) = \inf_{v \in U} \mathcal{L}(v, q). \quad (10.53)$$

We call the primal problem the minimization problem

$$\inf_{v \in U} \mathcal{J}(v), \quad (10.54)$$

and the dual problem the maximization problem

$$\sup_{q \in P} \mathcal{G}(q). \quad (10.55)$$

**Remark 10.3.7** Of course, without supplementary hypotheses, it can happen that  $\mathcal{J}(v) = +\infty$  for certain values of  $v$  or that  $\mathcal{G}(q) = -\infty$  for certain values of  $q$ . But the assumed existence of the saddle point  $(u, p)$  in definition 10.3.6 assures us that the **domains** of  $\mathcal{J}$  and  $\mathcal{G}$  (that is, the sets  $\{v \in U, \mathcal{J}(v) < +\infty\}$  and  $\{q \in P, \mathcal{G}(q) > -\infty\}$  over which these functions are well defined) are not empty, since (10.52) shows that  $\mathcal{J}(u) = \mathcal{G}(p) = \mathcal{L}(u, p)$ . The primal and dual problems therefore have a meaning. The following result shows that these two problems are closely linked to the saddle point  $(u, p)$ . •

**Theorem 10.3.8 (duality)** *The couple  $(u, p)$  is a saddle point of  $\mathcal{L}$  over  $U \times P$  if and only if*

$$\mathcal{J}(u) = \min_{v \in U} \mathcal{J}(v) = \max_{q \in P} \mathcal{G}(q) = \mathcal{G}(p). \quad (10.56)$$

**Remark 10.3.9** By the definition (10.53) of  $\mathcal{J}$  and  $\mathcal{G}$ , (10.56) is equivalent to

$$\mathcal{J}(u) = \min_{v \in U} \left( \sup_{q \in P} \mathcal{L}(v, q) \right) = \max_{q \in P} \left( \inf_{v \in U} \mathcal{L}(v, q) \right) = \mathcal{G}(p). \quad (10.57)$$

If the sup and the inf are attained in (10.57) (that is to say that we can write max and min, respectively), we then see that (10.57) gives the possibility of changing the order of the min and of the max applied to the Lagrangian  $\mathcal{L}$ . This fact (which is false if  $\mathcal{L}$  does not have a saddle point) explains the name minimax which is often given to a saddle point. •

**Proof.** Let  $(u, p)$  be a saddle point of  $\mathcal{L}$  over  $U \times P$ . Let us denote by  $\mathcal{L}^* = \mathcal{L}(u, p)$ . For  $v \in U$ , it is clear from (10.53) that  $\mathcal{J}(v) \geq \mathcal{L}(v, p)$ , from which  $\mathcal{J}(v) \geq \mathcal{L}^*$  from (10.52). As  $\mathcal{J}(u) = \mathcal{L}^*$ , this shows that  $\mathcal{J}(u) = \inf_{v \in U} \mathcal{J}(v) = \mathcal{L}^*$ . We show in the same way that  $\mathcal{G}(p) = \sup_{q \in P} \mathcal{G}(q) = \mathcal{L}^*$ .

Conversely, let us assume that (10.56) holds and set  $\mathcal{L}^* = \mathcal{J}(u)$ . Definition (10.53) of  $\mathcal{J}$  shows that

$$\mathcal{L}(u, q) \leq \mathcal{J}(u) = \mathcal{L}^* \quad \forall q \in P. \quad (10.58)$$

Likewise, we also have:

$$\mathcal{L}(v, p) \geq \mathcal{G}(p) = \mathcal{L}^* \quad \forall v \in U, \quad (10.59)$$

and we deduce easily from (10.58) to (10.59) that  $\mathcal{L}(u, p) = \mathcal{L}^*$ , which shows that  $(u, p)$  is a saddle point.  $\square$

**Remark 10.3.10** Likewise if the Lagrangian  $\mathcal{L}$  does not have a saddle point over  $U \times P$ , we still have the following elementary inequality, called **weak duality**

$$\inf_{v \in U} \left( \sup_{q \in P} \mathcal{L}(v, q) \right) \geq \sup_{q \in P} \left( \inf_{v \in U} \mathcal{L}(v, q) \right). \quad (10.60)$$

In effect, for all  $v \in U$  and  $q \in P$ ,  $\mathcal{L}(v, q) \geq \inf_{v' \in U} \mathcal{L}(v', q)$ , therefore  $\sup_{q \in P} \mathcal{L}(v, q) \geq \sup_{q \in P} \inf_{v' \in U} \mathcal{L}(v', q)$ , and since this is true for all  $v \in U$ ,  $\inf_{v \in U} \sup_{q \in P} \mathcal{L}(v, q) \geq \sup_{q \in P} \inf_{v' \in U} \mathcal{L}(v', q)$ , which gives (10.60). The (positive) difference between the two members of the inequality (10.60) is called the **duality gap**.  $\bullet$

**Exercise 10.3.2** Give an example of a Lagrangian for which the inequality (10.60) is strict with its two members finite.

**Exercise 10.3.3** Let  $U$  (respectively  $P$ ) be a convex compact nonempty subset of  $V$  (respectively  $Q$ ). We assume that the Lagrangian is such that  $v \rightarrow \mathcal{L}(v, q)$  is convex over  $U$  for all  $q \in P$ , and  $q \rightarrow \mathcal{L}(v, q)$  is concave over  $P$  for all  $v \in U$ . Show then the existence of a saddle point of  $\mathcal{L}$  over  $U \times P$ .

## Application

We apply this duality result to the preceding problem of convex minimization with convex inequality constraints.

$$\inf_{v \in V, F(v) \leq 0} J(v) \quad (10.61)$$

with  $J$  and  $F = (F_1, \dots, F_M)$  convex over  $V$ . We introduce the Lagrangian

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v) \quad \forall (v, q) \in V \times (\mathbb{R}_+)^M.$$

In this framework, we easily see that, for all  $v \in V$ ,

$$\mathcal{J}(v) = \sup_{q \in (\mathbb{R}_+)^M} \mathcal{L}(v, q) = \begin{cases} J(v) & \text{if } F(v) \leq 0 \\ +\infty & \text{otherwise,} \end{cases} \quad (10.62)$$

which shows that the primal problem  $\inf_{v \in V} \mathcal{J}(v)$  is exactly the original problem (10.61)! On the other hand, the function  $\mathcal{G}(q)$  of the dual problem is well defined by (10.53), as (10.53)

is here a convex minimization problem. Further,  $\mathcal{G}(q)$  is a concave function since it is the infimum of affine functions (see exercise 9.2.3). Consequently, the dual problem

$$\sup_{q \in (\mathbb{R}_+)^M} \mathcal{G}(q),$$

is a **simpler** concave maximization problem than the primal problem (10.61) since the constraints are linear! This characteristic is notably exploited in some numerical algorithms (cf. Uzawa's algorithm). A simple combination of the Kuhn–Tucker 10.3.4 and duality 10.3.8 theorems gives us the following result.

**Corollary 10.3.11** *We assume that the functions  $J, F_1, \dots, F_M$  are convex and differentiable over  $V$ . Take  $u \in V$  such that  $F(u) \leq 0$  and the constraints are qualified at  $u$  in the sense of definition 10.2.13. Then, if  $u$  is a global minimum of  $\mathcal{J}$  over  $V$ , there exists  $p \in (\mathbb{R}_+)^M$  such that*

1.  $p$  is a global maximum of  $\mathcal{G}$  over  $(\mathbb{R}_+)^M$ ,
2.  $(u, p)$  is a saddle point of the Lagrangian  $\mathcal{L}$  over  $V \times (\mathbb{R}_+)^M$ ,
3.  $(u, p) \in V \times (\mathbb{R}_+)^M$  satisfies the necessary and sufficient optimality condition

$$F(u) \leq 0, \quad p \geq 0, \quad p \cdot F(u) = 0, \quad J'(u) + p \cdot F'(u) = 0. \quad (10.63)$$

The most current application of corollary 10.3.11 is the following. Let us suppose that the dual problem of maximization is easier to solve than the primal problem (this is the case in general since its constraints are more simple). Then to calculate the solution  $u$  of the primal problem we proceed in two steps. First, we calculate the solution  $p$  of the dual problem. Second, we say that  $(u, p)$  is a saddle point of the Lagrangian, that is to say that we calculate  $u$ , the solution of the minimization problem **without constraint**

$$\min_{v \in V} \mathcal{L}(v, p).$$

Let us make precise that with the hypotheses made there is no *a priori* uniqueness of the solutions for all these problems. Let us also make precise that to obtain the existence of the minimum  $u$  in corollary 10.3.11 it is enough to add a hypothesis of strong convexity or of infinite behaviour at infinity on  $J$ .

**Remark 10.3.12** To illustrate corollary 10.3.11 and the interest in duality, we consider a quadratic minimization problem in  $\mathbb{R}^N$  with affine inequality constraints

$$\min_{v \in \mathbb{R}^N, \quad F(v) = Bv - c \leq 0} \left\{ J(v) = \frac{1}{2} Av \cdot v - b \cdot v \right\}, \quad (10.64)$$

where  $A$  is an  $N \times N$  symmetric positive definite matrix,  $b \in \mathbb{R}^N$ ,  $B$  an  $M \times N$  matrix, and  $c \in \mathbb{R}^M$ . The Lagrangian is given by

$$\mathcal{L}(v, q) = \frac{1}{2} Av \cdot v - b \cdot v + q \cdot (Bv - c) \quad \forall (v, q) \in \mathbb{R}^N \times (\mathbb{R}_+)^M. \quad (10.65)$$

We have already carried out, in (10.62), the calculation of  $\mathcal{J}$ , and said that the primal problem is exactly (10.64). Let us now examine the dual problem. For  $q \in (\mathbb{R}_+)^M$ , the problem

$$\min_{v \in \mathbb{R}^N} \mathcal{L}(v, q)$$

has a unique solution since  $v \rightarrow \mathcal{L}(v, q)$  is a strongly convex function. This solution satisfies  $\frac{\partial \mathcal{L}}{\partial v}(v, q) = Av - b + B^*q = 0$ , let  $v = A^{-1}(b - B^*q)$ . We therefore obtain

$$\mathcal{G}(q) = \mathcal{L}(A^{-1}(b - B^*q), q),$$

and the dual problem is finally written

$$\sup_{q \geq 0} \left( -\frac{1}{2}q \cdot BA^{-1}B^*q + (BA^{-1}b - c) \cdot q - \frac{1}{2}A^{-1}b \cdot b \right). \quad (10.66)$$

Admittedly, the functional to be maximized in (10.66) does not have a particularly sympathetic allure. It is again a problem with quadratic functional and affine constraints. However, corollary 10.3.11 assures us that it has a solution. We can see besides that this solution is not inevitably unique (except if the matrix  $B$  has range  $M$  since the matrix  $BA^{-1}B^*$  is then positive definite). But the important advantage of the dual problem (10.66) comes from the fact that the constraints ( $q \geq 0$ ) are expressed in a particularly simple form, simpler than for the primal problem; and we see in Section 10.5.3 that this advantage can be used to develop a computational algorithm for the solution of the primal problem. •

Let us finish with an entertaining exercise which shows the relation between saddle point or minimax problems and game theory.

**Exercise 10.3.4** Take a rectangular matrix

$$A = \begin{pmatrix} 1 & 0 & 4 & 2 & 3 & 5 \\ -3 & 2 & -1 & 2 & -5 & 2 \\ -4 & 2 & -2 & 0 & -1 & 2 \\ -2 & 4 & -1 & 6 & -2 & 2 \\ -1 & 2 & -6 & 3 & -1 & 1 \end{pmatrix}.$$

We assume that one of the two players chooses a line  $i$ , the other a column  $j$ , without one knowing the choice of the other. When their choices are revealed, the gain (or the loss, depending on the sign) of the first player is determined by the coefficient  $a_{ij}$  of the matrix  $A$  (the other player receiving or paying  $-a_{ij}$ ). Show that the optimal strategy to minimize the risk leads to a minimax problem that we shall solve. Is the game fair for this matrix  $A$ ?

## 10.4 Applications

In this section we shall study some applications of the results of the preceding sections. Let us point out that another application, linear programming, will be treated in the next chapter because of its importance in operations research.

### 10.4.1 Dual or complementary energy

In Chapter 5 we have seen that the solution of the following boundary value problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (10.67)$$

where  $\Omega$  is a bounded open set of  $\mathbb{R}^N$  and  $f \in L^2(\Omega)$ , is equivalent to the minimization of an energy

$$\min_{v \in H_0^1(\Omega)} \left\{ J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx \right\} \quad (10.68)$$

(see proposition 5.2.7). We have seen in exercise 9.2.8 that (10.68) has a unique minimum and, in exercise 10.2.4, that its Euler equation is the variational formulation of (10.67). The physical significance of the energy (10.68) is obvious. For example, if (10.67) models the deformation of an elastic membrane ( $u$  is the normal displacement under the action of forces  $f$ ), the solution is the displacement which minimizes the sum of the elastic energy of deformation and of the potential energy of the exterior forces. We propose to show that the duality theory allows us to associate with (10.67) a **second minimization principle** bringing into play an energy, called a **complementary** (or dual) energy in mechanics, whose physical significance is just as important as that of (10.68).

We shall introduce a Lagrangian associated with the primal energy (10.68) although this one does not have constraints. To do this we introduce an intermediate variable  $e \in L^2(\Omega)^N$  and a constraint  $e = \nabla v$ . Then (10.68) is equivalent to

$$\min_{\substack{v \in H_0^1(\Omega), \\ e \in L^2(\Omega)^N \\ e = \nabla v}} \left\{ \tilde{J}(v, e) = \frac{1}{2} \int_{\Omega} |e|^2 dx - \int_{\Omega} f v dx \right\}.$$

We introduce an intermediate Lagrangian for this problem

$$\mathcal{M}(e, v, \tau) = \tilde{J}(v, e) + \int_{\Omega} \tau \cdot (\nabla v - e) dx,$$

with a Lagrange multiplier  $\tau \in L^2(\Omega)^N$ . We now eliminate  $e$  to obtain the Lagrangian sought

$$\mathcal{L}(v, \tau) = \min_{e \in L^2(\Omega)^N} \mathcal{M}(e, v, \tau).$$

As  $e \rightarrow \mathcal{M}(e, v, \tau)$  is strongly convex, there exists a unique minimum point, and an easy calculation shows that

$$\mathcal{L}(v, \tau) = -\frac{1}{2} \int_{\Omega} |\tau|^2 dx - \int_{\Omega} f v dx + \int_{\Omega} \tau \cdot \nabla v dx. \quad (10.69)$$

We easily see that the primal problem associated with the Lagrangian (10.69) is (10.68)

$$\left( \max_{\tau \in L^2(\Omega)^N} \mathcal{L}(v, \tau) \right) = J(v),$$

and that the dual problem is

$$\left( \min_{v \in H_0^1(\Omega)} \mathcal{L}(v, \tau) \right) = G(\tau) = \begin{cases} -\frac{1}{2} \int_{\Omega} |\tau|^2 dx & \text{if } -\operatorname{div} \tau = f \text{ in } \Omega \\ -\infty & \text{otherwise.} \end{cases} \quad (10.70)$$

We can now state the principal result.

**Theorem 10.4.1** *There exists a unique saddle point  $(u, \sigma)$  of the Lagrangian  $\mathcal{L}(v, \tau)$  over  $H_0^1(\Omega) \times L^2(\Omega)^N$*

$$\mathcal{L}(u, \sigma) = \max_{\tau \in L^2(\Omega)^N} \min_{v \in H_0^1(\Omega)} \mathcal{L}(v, \tau) = \min_{v \in H_0^1(\Omega)} \max_{\tau \in L^2(\Omega)^N} \mathcal{L}(v, \tau).$$

*In other words,  $u$  is the unique minimum point of  $J(v)$  in  $H_0^1(\Omega)$ ,  $\sigma$  is the unique maximum point of  $G(\tau)$  in  $L^2(\Omega)^N$ ,*

$$J(u) = \min_{v \in H_0^1(\Omega)} J(v) = \max_{\tau \in L^2(\Omega)^N} G(\tau) = G(\sigma),$$

*and they are linked by the relation  $\sigma = \nabla u$ .*

**Remark 10.4.2** The dual problem (10.70) has a clear physical interpretation. As  $\max G(\tau) = -\min(-G(\tau))$ , it minimizes the (complementary) energy of mechanical constraints  $\frac{1}{2} \int_{\Omega} |\tau|^2 dx$  in the set of **statically admissible** stress fields, that is to say satisfying the equilibrium of forces  $-\operatorname{div} \tau = f$  in  $\Omega$ . In this dual formulation, the displacement  $v$  appears as the Lagrange multiplier of the equilibrium constraint  $-\operatorname{div} \tau = f$ . Another consequence of theorem 10.4.1 is that we always have  $G(\tau) \leq J(v)$  which allows us to obtain bounds over the primal or dual energies. Let us remark that we also have

$$J(u) = G(\sigma) = \frac{1}{2} \int_{\Omega} f u \, dx$$

which is none other than half of the work of the exterior forces. •

**Proof.** The proof could be an immediate consequence of corollary 10.3.11 (by remarking that an affine equality constraint is written as two opposed affine inequalities) if this theorem was not restricted (here) to a finite number of constraints. Now, in the dual problem (10.70) there are an infinite number of constraints since the constraint  $-\operatorname{div} \tau = f$  holds for almost all points  $x \in \Omega$ . Nevertheless, the result is true and it is easy to see why. In effect, by construction we have

$$G(\tau) \leq \mathcal{L}(v, \tau) \leq J(v),$$

and we know that the primal and dual problems have a unique minimum point  $u$  and  $\sigma$ , respectively. Now, as  $u$  is a solution of (10.67), a simple integration by parts shows that

$$J(u) = \frac{1}{2} \int_{\Omega} |\nabla u|^2 dx - \int_{\Omega} f u \, dx = -\frac{1}{2} \int_{\Omega} |\nabla u|^2 dx = -\frac{1}{2} \int_{\Omega} f u \, dx.$$

If we define  $\sigma = \nabla u$ , we deduce from (10.67) that  $-\operatorname{div} \sigma = f$ , and we therefore obtain

$$G(\tau) \leq J(u) = G(\sigma),$$

that is to say that  $\sigma$  is the maximum point of  $G$ , therefore  $(u, \sigma)$  is the saddle point of  $\mathcal{L}(v, \tau)$ . □

## 10.4.2 Optimal command

Here we solve example 9.1.9 of a problem of optimal command called the linear-quadratic system. We consider the linear differential system whose unknown (the state of the system)  $y(t)$  has values in  $\mathbb{R}^N$

$$\begin{cases} \frac{dy}{dt} = Ay + Bv + f & \text{for } 0 \leq t \leq T \\ y(0) = y_0 \end{cases} \quad (10.71)$$

where  $y_0 \in \mathbb{R}^N$  is the initial state of the system,  $f(t) \in \mathbb{R}^N$  is a source term,  $v(t) \in \mathbb{R}^M$  is the command that allows us to act over the system, and  $A$  and  $B$  are two constant matrices of dimensions  $N \times N$  and  $N \times M$ , respectively.

We want to choose the command  $v$  to minimize a quadratic criterion

$$J(v) = \frac{1}{2} \int_0^T Rv(t) \cdot v(t) dt + \frac{1}{2} \int_0^T Q(y - z)(t) \cdot (y - z)(t) dt + \frac{1}{2} D(y(T) - z_T) \cdot (y(T) - z_T),$$

where  $z(t)$  is a ‘target’ trajectory,  $z_T$  is a final ‘target’ position, and  $R, Q, D$  three symmetric positive matrices one of which namely  $R$  is assumed positive definite. Let us remark that the function  $y(t)$  depends on the variable  $v$  through (10.71).

To be able to apply the preceding optimization results, we choose to look for  $v$  in the Hilbert space  $L^2(]0, T[; \mathbb{R}^M)$  of functions of  $]0, T[$  in  $\mathbb{R}^M$  which are square integrable. (The ‘more natural’ space of continuous functions is unfortunately not a Hilbert space.) To take account of possible constraints on the command, we introduce a convex closed nonempty set  $K$  of  $\mathbb{R}^M$  which represents the set of admissible commands. The minimization problem is therefore

$$\inf_{v(t) \in L^2(]0, T[; K)} J(v). \quad (10.72)$$

Let us start by verifying that the system (10.71) is well-posed.

**Lemma 10.4.3** *We assume that  $f(t) \in L^2(]0, T[; \mathbb{R}^N)$  and  $v(t) \in L^2(]0, T[; K)$ . Then (10.71) has a unique solution  $y(t) \in H^1(]0, T[; \mathbb{R}^N)$  which is, moreover, continuous over  $[0, T]$ .*

**Proof.** This existence and uniqueness result is well known if  $f$  and  $v$  are continuous. It is no more difficult in the framework of  $L^2$ . We use the explicit representation formula of the solution

$$y(t) = \exp(tA)y_0 + \int_0^t \exp((t-s)A)(Bv + f)(s) ds$$

which allows us to verify the existence and uniqueness of  $y$  in  $H^1(]0, T[; \mathbb{R}^N)$ . Lemma 4.3.3 tells us finally that  $y$  is continuous over  $[0, T]$ .  $\square$

We can then show the existence and uniqueness of the optimal command.

**Proposition 10.4.4** *There exists a unique  $u \in L^2(]0, T[; K)$  which minimizes (10.72). This optimal command  $u$  is characterized by*

$$\begin{aligned} \int_0^T Q(y_u - z) \cdot (y_v - y_u) dt + \int_0^T Ru \cdot (v - u) dt \\ + D(y_u(T) - z_T) \cdot (y_v(T) - y_u(T)) \geq 0, \end{aligned} \quad (10.73)$$

for all  $v \in L^2([0, T]; K)$ , where  $y_v$  denotes the solution of (10.71) associated with the command  $v$ .

**Proof.** We start by remarking that  $v \rightarrow y$  is an affine function. In effect, by the linearity of (10.71) we have  $y_v = \tilde{y}_v + \hat{y}$ , where  $\tilde{y}_v$  is a solution of

$$\begin{cases} \frac{d\tilde{y}_v}{dt} = A\tilde{y}_v + Bv & \text{for } 0 \leq t \leq T \\ \tilde{y}_v(0) = 0 \end{cases} \quad (10.74)$$

and  $\hat{y}$  is a solution of

$$\begin{cases} \frac{d\hat{y}}{dt} = A\hat{y} + f & \text{for } 0 \leq t \leq T \\ \hat{y}(0) = y_0 \end{cases}$$

It is clear that  $\hat{y}$  does not depend on  $v$  and  $v \rightarrow \tilde{y}_v$  is continuous and linear from  $L^2([0, T]; K)$  into  $H^1([0, T]; \mathbb{R}^N)$ . Consequently,  $v \rightarrow J(v)$  is a positive quadratic function of  $v$  (more precisely, the sum of a quadratic form and an affine function), therefore  $J$  is convex, and also strongly convex since the matrix  $R$  is positive definite. As  $L^2([0, T]; K)$  is a convex closed nonempty set, theorem 9.2.6 allows us to conclude the existence and uniqueness of the minimum point  $u$  of (10.72). On the other hand, the necessary and sufficient optimality condition of theorem 10.2.1 is  $\langle J'(u), v - u \rangle \geq 0$ . To calculate the gradient, the surest and simplest method is to calculate

$$\lim_{\epsilon \rightarrow 0} \frac{J(u + \epsilon w) - J(u)}{\epsilon} = \langle J'(u), w \rangle.$$

As  $J(v)$  is quadratic the calculation is very simple since  $y_{u+\epsilon w} = y_u + \epsilon \tilde{y}_w$ . We easily obtain (10.73) by remarking that  $y_u - y_v = \tilde{y}_u - \tilde{y}_v$ .  $\square$

**Remark 10.4.5** By making explicit the optimality condition of (10.72) we have in fact calculated the gradient  $J'(w)$  for all  $w \in L^2[0, T[$  (and not only for the minimum  $u$ ), which is useful for numerical methods for minimization (see section 10.5). We have obtained

$$\int_0^T J'(w)v \, dt = \int_0^T R w \cdot v \, dt + \int_0^T Q(y_w - z) \cdot \tilde{y}_v \, dt + D(y_w(T) - z_T) \cdot \tilde{y}_v(T), \quad (10.75)$$

where  $v$  is an arbitrary function of  $L^2[0, T[$ .  $\bullet$

The necessary and sufficient optimality condition (10.73) is in fact not exploitable! In effect, to test the optimality of  $u$  it is necessary for each test function  $v$  to calculate the corresponding state  $y_v$ . Another way to see this difficulty is the impossibility of obtaining an explicit expression for  $J'(u)$  starting from (10.75). To circumvent this difficulty we have recourse to the idea of an **adjoint state** which is one of the most profound ideas in the theory of optimal control. Let us show how to proceed using the example studied in this subsection (we shall give the general idea in remark 10.4.8 below). For the problem (10.72) we defined the adjoint state  $p$  as the unique solution of

$$\begin{cases} \frac{dp}{dt} = -A^*p - Q(y - z) & \text{for } 0 \leq t \leq T \\ p(T) = D(y(T) - z_T) \end{cases} \quad (10.76)$$



where  $y$  is the solution of (10.71) for the command  $u$ . The name adjoint state comes from the fact that it is the adjoint matrix  $A^*$  which appears in (10.76). The interest in the adjoint state is that it allows us to obtain an explicit expression for  $J'(u)$ .

**Theorem 10.4.6** *The derivative of  $J$  at  $u$  is given by*

$$J'(u) = B^*p + Ru. \quad (10.77)$$

*In particular, the necessary and sufficient optimality condition of the problem (10.72) is*

$$\int_0^T (B^*p + Ru) \cdot (v - u) dt \geq 0 \quad \forall v \in L^2([0, T]; K). \quad (10.78)$$

**Remark 10.4.7** Formula (10.77) generalizes for all  $w \in L^2[0, T]$  in  $J'(w) = B^*p + Rw$ , even if it means calculating  $p$  by (10.76) by using the state  $y$  corresponding to the command  $w$ . Theorem 10.4.6 gives an explicit expression for the gradient at the cost of the supplementary solution of the adjoint system (10.76). This is a fundamental difference with the formula (10.75) which, for each test function  $v$ , needs the solution of the system (10.71) with the command  $v$ . •

**Proof.** Let  $p$  be the solution of (10.76) and  $\tilde{y}_v$  that of (10.74). The idea is to multiply (10.76) by  $\tilde{y}_v$  and (10.74) by  $p$ , to integrate by parts and to compare the results. More precisely, we calculate the following quantity in two different ways. First, by integrating and by taking account of the initial conditions  $\tilde{y}_v(0) = 0$  and  $p(T) = D(y(T) - z_T)$ , we have

$$\int_0^T \left( \frac{dp}{dt} \cdot \tilde{y}_v + p \cdot \frac{d\tilde{y}_v}{dt} \right) dt = D(y(T) - z_T) \cdot \tilde{y}_v(T). \quad (10.79)$$

On the other hand, by using the equations we obtain

$$\int_0^T \left( \frac{dp}{dt} \cdot \tilde{y}_v + p \cdot \frac{d\tilde{y}_v}{dt} \right) dt = - \int_0^T Q(y - z) \cdot \tilde{y}_v dt + \int_0^T Bv \cdot p dt. \quad (10.80)$$

We deduce, from the equality between (10.79) and (10.80), a simplification of the expression (10.75) of the derivative

$$\int_0^T J'(u)v dt = \int_0^T Ru \cdot v dt + \int_0^T Bv \cdot p dt,$$

which gives the results (10.77) and (10.78). □

**Remark 10.4.8** How can we **guess** the problem (10.76) which defines the adjoint state in order to simplify the expression of  $J'(v)$ ? Once again, the main idea is the introduction of a Lagrangian associated with the minimization problem (10.72). We consider the equation of state (10.71) as a constraint between two independent variables  $v$  and  $y$  and we define the Lagrangian as the sum of  $J(v)$  and of the equation of state multiplied by  $p$ , that is to say

$$\begin{aligned} \mathcal{L}(v, y, p) &= \int_0^T Rv(t) \cdot v(t) dt + \int_0^T Q(y - z)(t) \cdot (y - z)(t) dt \\ &\quad + D(y(T) - z_T) \cdot (y(T) - z_T) + \int_0^T p \cdot \left( -\frac{dy}{dt} + Ay + Bv + f \right) dt \\ &\quad - p(0) \cdot (y(0) - y_0), \end{aligned}$$

where  $p$  is the Lagrange multiplier for the constraint (10.71) between  $v$  and  $y$ . Formally, the optimality conditions of (10.72) are obtained by saying that the Lagrangian is stationary, that is to say that

$$\frac{\partial \mathcal{L}}{\partial v} = \frac{\partial \mathcal{L}}{\partial y} = \frac{\partial \mathcal{L}}{\partial p} = 0.$$

The first derivative gives the optimality condition (10.77), the second gives the equation satisfied by the adjoint state  $p$ , and the third the equation satisfied by the state  $y$ . Let us insist on the fact that this calculation is purely formal, but that in general it gives the ‘good’ equation of the adjoint state. •

Starting from theorem 10.4.6 we can obtain some qualitative properties of the solution  $y$  and of the optimal command  $u$  (see exercise 10.4.1), and construct a numerical method to minimize (10.72) by a gradient type algorithm. In the absence of constraints on the command, that is to say if  $K = \mathbb{R}^M$ , we can even go further in the analysis and find a ‘law of command’ which gives the adjoint state  $p$  (and therefore the optimal command  $u = -R^{-1}B^*p$  because of (10.78)).

**Proposition 10.4.9** *We assume that  $K = \mathbb{R}^M$ ,  $f = 0$ ,  $z = 0$ , and  $z_T = 0$ . We choose the optimal command  $u = -R^{-1}B^*P$ . Let  $P(t)$  be the matrix valued (of order  $N$ ) function of  $[0, T]$  which is the unique solution of*

$$\begin{cases} \frac{dP}{dt} = -A^*P - PA + PBR^{-1}B^*P - Q & \text{for } 0 \leq t \leq T \\ P(T) = D \end{cases} \quad (10.81)$$

*Then  $P(t)$  is symmetric positive for all  $t \in [0, T]$  and we have  $p(t) = P(t)y(t)$ .*

**Proof.** For all  $t \in [0, T]$ , the mapping which, for  $y_0 \in \mathbb{R}^N$ , gives  $(y, p)(t)$  is clearly linear. It is injective from exercise 10.4.1. Consequently, as  $y_0$  varies the relation between  $y(t)$  and  $p(t)$  is linear, and there exists a matrix  $P(t)$  of order  $N$  such that  $p(t) = P(t)y(t)$ . By differentiating this expression and by using the equations (10.71) and (10.76), we obtain

$$\frac{dP}{dt}y = -A^*Py - PAy + PBR^{-1}B^*Py - Qy$$

for all  $y(t)$  (which is arbitrary since  $y_0$  is). We deduce the equation in (10.81). We obtain the final condition  $P(T) = D$  since  $p(T) = Dy(T)$  and  $y(T)$  is arbitrary in  $\mathbb{R}^N$ . We have the uniqueness of the solution of (10.81). □

**Exercise 10.4.1** We assume that  $K = \mathbb{R}^M$ ,  $f = 0$ ,  $z = 0$ , and  $z_T = 0$ . Show that, for all  $t \in [0, T]$ , the optimal solution satisfies

$$p(t) \cdot y(t) = Dy(T) \cdot y(T) + \int_t^T Qy(s) \cdot y(s) ds + \int_t^T R^{-1}B^*p(s) \cdot B^*p(s) ds.$$

Deduce that if there exists  $t_0 \in [0, T]$  such that  $y(t_0) = 0$ , then  $y(t) = p(t) = 0$  for all  $t \in [0, T]$ . Interpret this result.

**Exercise 10.4.2** Obtain the equivalent of proposition 10.4.4 and of theorem 10.4.6 for the parabolic system

$$\begin{cases} \frac{\partial y}{\partial t} - \Delta y = v + f & \text{in } ]0, T[ \times \Omega \\ y = 0 & \text{on } ]0, T[ \times \partial\Omega \\ y(0) = y_0 & \text{in } \Omega \end{cases}$$

where  $y_0 \in L^2(\Omega)$ ,  $f \in L^2(]0, T[ \times \Omega)$ ,  $v \in L^2(]0, T[ \times \Omega)$  is the command, and we minimize

$$\inf_{v \in L^2(]0, T[ \times \Omega)} J(v) = \int_0^T \int_{\Omega} v^2 dt dx + \int_0^T \int_{\Omega} |y - z|^2 dt dx + \int_{\Omega} |y(T) - z_T|^2 dx,$$

where  $z \in L^2(]0, T[ \times \Omega)$  and  $z_T \in L^2(\Omega)$ .

**Exercise 10.4.3** Generalize the preceding exercise to the wave equation.

### 10.4.3 Optimization of distributed systems

We solve here an example 9.1.12 of the control of an elastic membrane deformed by an exterior force  $f$  and which has a fixed shape. The behaviour of the membrane is modelled by

$$\begin{cases} -\Delta u = f + v & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega, \end{cases} \quad (10.82)$$

where  $u$  is the vertical displacement of the membrane and  $v$  is a control force which will be the optimization variable. We are given an open set  $\omega \subset \Omega$  over which the control acts and two limiting functions  $v_{\min} \leq v_{\max}$  in  $L^2(\omega)$ . In all that follows we assume that the functions of  $L^2(\omega)$  are extended by zero in  $\Omega \setminus \omega$ . We then define the set of admissible controls

$$K = \{v \in L^2(\omega) \text{ such that } v_{\min}(x) \leq v(x) \leq v_{\max}(x) \text{ in } \omega \text{ and } v = 0 \text{ in } \Omega \setminus \omega\}. \quad (10.83)$$

If  $f \in L^2(\Omega)$ , theorem 5.2.2 tells us that there exists a unique solution  $u \in H_0^1(\Omega)$ . We want to control the membrane so that it adopts a displacement  $u_0 \in L^2(\Omega)$ . We define a cost function

$$J(v) = \frac{1}{2} \int_{\Omega} (|u - u_0|^2 + c|v|^2) dx, \quad (10.84)$$

where  $u$  is the solution of (10.82) (and therefore depends on  $v$ ) and  $c > 0$ . The optimization problem is written

$$\inf_{v \in K} J(v). \quad (10.85)$$

**Proposition 10.4.10** *There exists a unique optimal control  $\bar{v} \in K$  for the problem (10.85).*

**Proof.** We remark that the function  $v \rightarrow u$  is affine from  $L^2(\Omega)$  into  $H_0^1(\Omega)$ . Consequently,  $J(v)$  is a positive quadratic function of  $v$ , therefore it is convex. It is likewise strongly convex since  $J(v) \geq c\|v\|_{L^2(\Omega)}^2$ . On the other hand,  $K$  is a convex closed nonempty set of  $L^2(\Omega)$ . Consequently, theorem 9.2.6 allows us to conclude the existence and uniqueness of the minimum point of (10.85).  $\square$

To obtain a necessary optimality condition which is exploitable, we introduce, as in Section 10.4.2, an adjoint state  $p$  defined as the unique solution in  $H_0^1(\Omega)$  of

$$\begin{cases} -\Delta p = u - u_0 & \text{in } \Omega \\ p = 0 & \text{on } \partial\Omega. \end{cases} \quad (10.86)$$

**Proposition 10.4.11** *The cost function  $J(v)$  is differentiable over  $K$  and we have*

$$J'(v) = p + cv,$$

where  $p$  (which depends on  $v$ ) is given by (10.86). Consequently, the necessary and sufficient optimality condition for the optimal control  $\bar{v}$  is

$$-\Delta \bar{u} = f + \bar{v} \quad \text{in } \Omega, \quad \bar{u} \in H_0^1(\Omega), \quad (10.87)$$

$$-\Delta \bar{p} = \bar{u} - u_0 \quad \text{in } \Omega, \quad \bar{p} \in H_0^1(\Omega), \quad (10.88)$$

$$\bar{v} = \mathbb{I}_\omega P_{[v_{\min}(x), v_{\max}(x)]} \left( -\frac{\bar{p}}{c} \right), \quad (10.89)$$

where  $\mathbb{I}_\omega$  is the characteristic function of  $\omega$  (that is to say which is 1 in  $\omega$  and 0 in  $\Omega \setminus \omega$ ) and  $P_{[v_{\min}(x), v_{\max}(x)]}$  is the orthogonal projection operator over the segment  $[v_{\min}(x), v_{\max}(x)]$  defined by  $P_{[v_{\min}(x), v_{\max}(x)]} w = \min(v_{\max}(x), \max(v_{\min}(x), w(x)))$ .

**Proof.** As in proposition 10.4.4, the simplest and surest method to calculate the gradient is

$$\lim_{\epsilon \rightarrow 0} \frac{J(v + \epsilon w) - J(v)}{\epsilon} = \int_{\Omega} J'(v) w \, dx.$$

As  $J(v)$  is quadratic the calculation is very simple and we obtain

$$\int_{\Omega} J'(v) w \, dx = \int_{\Omega} ((u - u_0) \tilde{u}_w + cvw) \, dx,$$

where  $\tilde{u}_w$  is given by

$$\begin{cases} -\Delta \tilde{u}_w = w & \text{in } \Omega \\ \tilde{u}_w = 0 & \text{on } \partial\Omega. \end{cases} \quad (10.90)$$

To simplify the expression of the gradient we use the adjoint state: we multiply (10.90) by  $p$  and (10.86) by  $\tilde{u}_w$  and we integrate by parts

$$\begin{aligned} \int_{\Omega} \nabla p \cdot \nabla \tilde{u}_w \, dx &= \int_{\Omega} (u - u_0) \tilde{u}_w \, dx \\ \int_{\Omega} \nabla \tilde{u}_w \cdot \nabla p \, dx &= \int_{\Omega} wp \, dx \end{aligned}$$

By comparison of these two equalities we deduce that

$$\int_{\Omega} J'(v) w \, dx = \int_{\Omega} (p + cv) w \, dx,$$

from which we obtain the expression for the gradient. The necessary and sufficient optimality condition given by theorem 10.2.1 is

$$\int_{\Omega} (\bar{p} + c\bar{v}) (w - \bar{v}) \, dx \geq 0 \quad \forall w \in K. \quad (10.91)$$

By taking  $w$  equal to  $\bar{v}$  everywhere except over a small ball in  $\omega$ , then letting the radius of this ball tend to zero, we can ‘localize’ (10.91) at (almost) every point  $x$  of  $\omega$

$$(\bar{p}(x) + c\bar{v}(x)) (w(x) - \bar{v}(x)) \geq 0 \quad \forall w(x) \in [v_{\min}(x), v_{\max}(x)].$$

This last condition is only the definition of  $\bar{v}(x)$  as the orthogonal projection of  $-\bar{p}(x)/c$  on the segment  $[v_{\min}(x), v_{\max}(x)]$  (see theorem 12.1.10). Finally, we obtain (10.89) by remarking that the support of the functions of  $K$  is restricted to  $\omega$ .  $\square$

**Remark 10.4.12** As in remark 10.4.8 we explain how to find the form of (10.86) which defines the adjoint state. We introduce the Lagrangian associated with the minimization problem (10.85) under the constraint that the equation of state (10.82) (which links the two independent variables  $v$  and  $u$ ) satisfies

$$\mathcal{L}(v, u, p) = \frac{1}{2} \int_{\Omega} (|u - u_0|^2 + c|v|^2) dx + \int_{\Omega} p(\Delta u + f + v) dx,$$

where  $p$  is the Lagrange multiplier for the constraint (10.82) between  $v$  and  $u$ . Formally, the optimality conditions are obtained by saying that the Lagrangian is stationary, that is to say that

$$\frac{\partial \mathcal{L}}{\partial v} = \frac{\partial \mathcal{L}}{\partial u} = \frac{\partial \mathcal{L}}{\partial p} = 0.$$

The first derivative gives the optimality condition (10.89), the second gives the equation satisfied by the adjoint state  $p$ , and the third the equation satisfied by the state  $u$ .  $\bullet$

## 10.5 Numerical algorithms

### 10.5.1 Introduction

The object of this section is to present and analyse some algorithms which allow us to calculate, or more exactly to **approximate** the solution of the optimization problems studied above. All the algorithms studied here are used effectively in practice to solve concrete optimization problems by computer.

These algorithms are also all of an iterative nature: starting from a given initial  $u^0$ , each method constructs a sequence  $(u^n)_{n \in \mathbb{N}}$  which we shall show converges, under certain hypotheses, to the solution  $u$  of the optimization problem considered. After having shown the **convergence of these algorithms** (that is to say, the convergence of the sequence  $(u^n)$  to  $u$  whatever the choice of the initial data  $u^0$ ), we shall also say a word about their rate of convergence.

In all of this section we assume that the objective function  $J$  to be minimized is  $\alpha$ -convex and differentiable. This hypothesis of  $\alpha$ -convexity is rather strong, but we see later that it is crucial for the proofs of convergence of the algorithms. The application of the algorithms presented here to the minimization of convex functions which are not strongly convex can present some small difficulties, without mentioning the **great** difficulties which appear when we want to approximate the minimum of a nonconvex function! Typically, these algorithms cannot converge and oscillate

between several minimum points, or worse they converge to a local minimum, very far from a global minimum (in the nonconvex case, cf. proposition 9.2.3).

**Remark 10.5.1** We limit ourselves to deterministic algorithms and we say nothing of stochastic algorithms (simulated annealing, genetic algorithms, etc.). Besides the fact that their analysis calls on probability theory (which we do not discuss in this course), their use is very different. To put it simply, let us say that deterministic algorithms are the most efficient for the minimization of convex functions, while stochastic algorithms allow us to approximate **global** (not only local) minima of nonconvex functions (for a higher cost, however, in practice). •

## 10.5.2 Gradient algorithms (case without constraints)

Let us start by studying the practical solution of optimization problems in the absence of constraints. Let  $J$  be a function which is  $\alpha$ -convex and differentiable defined over the real Hilbert space  $V$ , we consider the problem without constraint

$$\inf_{v \in V} J(v). \quad (10.92)$$

From theorem 9.2.6 there exists a unique solution  $u$ , characterized by remark 10.2.2 by the Euler equation

$$J'(u) = 0.$$

### Gradient algorithm with optimal step

The gradient algorithm consists of ‘moving’ from an iterate  $u^n$  by following the line of greatest slope associated with the cost function  $J(v)$ . The direction of descent corresponding to this line of greatest slope from  $u^n$  is given by the gradient  $J'(u^n)$ . In effect, if we look for  $u^{n+1}$  in the form

$$u^{n+1} = u^n - \mu^n w^n, \quad (10.93)$$

with  $\mu^n > 0$  small and  $w^n$  a unit vector in  $V$ , it is with the choice of direction  $w_n = J'(u^n)/\|J'(u^n)\|$  that we can hope to find the smallest value of  $J(u^{n+1})$  (in the absence of other information such as higher derivatives or previous iterates).

This simple remark leads us, among the methods (10.93) which are called ‘methods of descent’, to the **gradient algorithm with optimal step**, in which we solve a succession of minimization problems with one real variable (even if  $V$  is not finite dimensional). Starting from an arbitrary  $u^0$  in  $V$ , we construct the sequence  $(u^n)$  defined by

$$u^{n+1} = u^n - \mu^n J'(u^n), \quad (10.94)$$

where  $\mu^n \in \mathbb{R}$  is chosen at each step such that

$$J(u^{n+1}) = \inf_{\mu \in \mathbb{R}} J(u^n - \mu J'(u^n)). \quad (10.95)$$

This algorithm converges as the following result shows.

**Theorem 10.5.2** *We assume that  $J$  is  $\alpha$ -convex and differentiable and that  $J'$  is Lipschitzian over all bounded sets of  $V$ , that is to say that*

$$\forall M > 0, \exists C_M > 0, \quad \|v\| + \|w\| \leq M \Rightarrow \|J'(v) - J'(w)\| \leq C_M \|v - w\|. \quad (10.96)$$

*Then the gradient algorithm with optimal step converges: for any  $u^0$ , the sequence  $(u^n)$  defined by (10.94) and (10.95) converges to the solution  $u$  of (10.92).*

**Proof.** The function  $f(\mu) = J(u^n - \mu J'(u^n))$  is strongly convex and differentiable over  $\mathbb{R}$  (if  $J'(u^n) \neq 0$ ; otherwise, we have already converged,  $u^n = u$ !). The minimization problem (10.95) therefore has a unique solution, characterized by the condition  $f'(\mu) = 0$ , which is also written

$$\langle J'(u^{n+1}), J'(u^n) \rangle = 0. \quad (10.97)$$

This shows that two consecutive ‘directions of descent’ are orthogonal.

Since (10.97) implies that  $\langle J'(u^{n+1}), u^{n+1} - u^n \rangle = 0$ , we deduce from the  $\alpha$ -convexity of  $J$  that

$$J(u^n) - J(u^{n+1}) \geq \frac{\alpha}{2} \|u^n - u^{n+1}\|^2. \quad (10.98)$$

Since the sequence  $J(u^n)$  is decreasing and bounded below (by  $J(u)$ ), it converges and the inequality (10.98) shows that  $u^{n+1} - u^n$  tends to 0. On the other hand, the  $\alpha$ -convexity of  $J$  and the fact that the sequence  $J(u^n)$  is bounded show that the sequence  $(u^n)$  is bounded: there exists a constant  $M$  such that

$$\|u^n\| \leq M.$$

Writing (10.96) for  $v = u^n$  and  $w = u^{n+1}$  and using (10.97), we easily see that  $\|J'(u^n)\| \leq C_M \|u^{n+1} - u^n\|$ , which shows that  $J'(u^n)$  tends to 0. The  $\alpha$ -convexity of  $J$  then gives

$$\alpha \|u^n - u\|^2 \leq \langle J'(u^n) - J'(u), u^n - u \rangle = \langle J'(u^n), u^n - u \rangle \leq \|J'(u^n)\| \|u^n - u\|,$$

which implies  $\alpha \|u^n - u\| \leq \|J'(u^n)\|$ , from which we deduce the convergence of the algorithm.  $\square$

**Remark 10.5.3** It is useful to note the practical interest of the last inequality of this proof: besides the proof of convergence, it gives an easily calculable upper bound on the error  $u^n - u$ .  $\bullet$

### Gradient algorithm with fixed step

The gradient algorithm with fixed step consists simply of the construction of a sequence  $u^n$  defined by

$$u^{n+1} = u^n - \mu J'(u^n), \quad (10.99)$$

where  $\mu$  is a fixed positive parameter. This method is therefore simpler than the gradient algorithm with optimal step, since at each step we save the cost of calculating the solution of (10.95). The following result shows under what hypotheses we can choose the parameter  $\mu$  to ensure the convergence.

**Theorem 10.5.4** *We assume that  $J$  is  $\alpha$ -convex and differentiable and that  $J'$  is Lipschitzian over  $V$ , that is to say that there exists a constant  $C > 0$  such that*

$$\|J'(v) - J'(w)\| \leq C\|v - w\| \quad \forall v, w \in V. \quad (10.100)$$

*Then, if  $0 < \mu < 2\alpha/C^2$ , the gradient algorithm with fixed step converges: for any  $u^0$ , the sequence  $(u^n)$  defined by (10.95) converges to the solution  $u$  of (10.92).*

**Proof.** Let us set  $v^n = u^n - u$ . As  $J'(u) = 0$ , we have  $v^{n+1} = v^n - \mu(J'(u^n) - J'(u))$ , from which it becomes

$$\begin{aligned} \|v^{n+1}\|^2 &= \|v^n\|^2 - 2\mu\langle J'(u^n) - J'(u), u^n - u \rangle + \mu^2\|J'(u^n) - J'(u)\|^2 \\ &\leq (1 - 2\alpha\mu + C^2\mu^2) \|v^n\|^2, \end{aligned} \quad (10.101)$$

from (10.100) and the  $\alpha$ -convexity. If  $0 < \mu < 2\alpha/C^2$ , it is easy to see that  $1 - 2\alpha\mu + C^2\mu^2 \in ]0, 1[$ , and the convergence is deduced from (10.101).  $\square$

**Remark 10.5.5** A simple adaptation of the preceding proof, left to the reader as an exercise, allows us to show convergence by replacing (10.100) by the weaker hypothesis (10.96). We must also note that, for the gradient algorithm with fixed step, compared to the gradient algorithm with optimal step, the sequence  $J(u^n)$  is not necessarily monotone.  $\bullet$

**Remark 10.5.6** Numerous other descent algorithms of the type (10.93) exist that we have not described here. In particular, in this class of algorithms, we meet the conjugate gradient method in which the direction of descent  $w^n$  depends not only on the gradient  $J'(u^n)$  but also on the directions of descent used in the preceding iterations. We present this method in Section 13.1.5, for the particular case of a quadratic functional of the type  $\frac{1}{2}Ax \cdot x - b \cdot x$ .  $\bullet$

**Remark 10.5.7** How do we choose between the two gradient algorithms we have just seen, and more generally between the different methods of numerical minimization which exist? A first criterion is the cost of each iteration. For example, as we have said, each iteration of the gradient algorithm with fixed step is less expensive than



an iteration of gradient algorithm with optimal step. Obviously, if we start from the same iterate  $u^n$ , an iteration of the gradient algorithm with optimal step decreases the cost function more than an iteration of the gradient algorithm with fixed step. We arrive therefore at the second criterion, often more important, which is that of the **rate of convergence** of the algorithm, which fixes the number of iterations necessary to make the error  $\|u^n - u\|$  less than a tolerance  $\epsilon$  fixed *a priori*.

For example, the inequality (10.101) shows that the convergence of the gradient algorithm with fixed step is at least geometrical, since

$$\|u^n - u\| \leq \gamma^n \|u^0 - u\| \quad \text{with} \quad \gamma = \sqrt{1 - 2\alpha\mu + \mu^2 C^2}.$$

This remark leads, subject to a more careful analysis, to preferring the parameter  $\mu$  to be the median value  $\alpha/C^2$  in the interval  $]0, 2\alpha/C^2[$ , which minimizes  $\gamma$ . In fact, we can show that the convergence of the two algorithms studied above is effectively geometrical in certain particular cases (which means that the quantity  $\|u^n - u\|^{1/n}$  has a finite limit, strictly between 0 and 1, when  $n$  tends to  $+\infty$ ). •

**Exercise 10.5.1** For  $V = \mathbb{R}^2$  and  $J(x, y) = ax^2 + by^2$  with  $a, b > 0$ , show that the gradient algorithm with optimal step converges in one iteration if  $a = b$  or if  $x^0 y^0 = 0$ , and that the convergence is geometrical in the other cases. Study also the convergence of the gradient algorithm with fixed step: for which values of the parameter  $\mu$  do we have convergence, for what value is it the most rapid?

### 10.5.3 Gradient algorithms (case with constraints)

We now study the solution of optimization problems with constraints

$$\inf_{v \in K} J(v), \tag{10.102}$$

where  $J$  is a function which is  $\alpha$ -convex and differentiable defined over  $K$ , a convex closed nonempty subset of the real Hilbert space  $V$ . Theorem 9.2.6 ensures the existence and uniqueness of the solution  $u$  of (10.102), characterized from theorem 10.2.1 by the condition

$$\langle J'(u), v - u \rangle \geq 0 \quad \forall v \in K. \tag{10.103}$$

According to the algorithms studied below, we will sometimes need to state supplementary hypotheses over the set  $K$ .

#### Gradient algorithm with fixed step and projection

The gradient algorithm with fixed step adapts to problem (10.102) with constraints starting from the following remark. For all real  $\mu > 0$ , (10.103) is written

$$\langle u - (u - \mu J'(u)), v - u \rangle \geq 0 \quad \forall v \in K. \tag{10.104}$$

Let us denote by  $P_K$  the projection operator over the convex set  $K$ , defined in theorem 12.1.10 (see remark 12.1.11). Then, from this theorem, (10.104) is none other than the characterization of  $u$  as the orthogonal projection of  $u - \mu J'(u)$  over  $K$ . In other words,

$$u = P_K(u - \mu J'(u)) \quad \forall \mu > 0. \quad (10.105)$$

It is easy to see that (10.105) is in fact equivalent to (10.103), and therefore characterizes the solution  $u$  of (10.102). The **gradient algorithm with fixed step and projection** algorithm (or more simply projected gradient) is then defined by the iteration

$$u^{n+1} = P_K(u^n - \mu J'(u^n)), \quad (10.106)$$

where  $\mu$  is a fixed positive parameter.

**Theorem 10.5.8** *We assume that  $J$  is  $\alpha$ -convex, differentiable and that  $J'$  is Lipschitzian over  $V$  (of constant  $C$ , see (10.100)). Then, if  $0 < \mu < 2\alpha/C^2$ , the gradient algorithm with fixed step and projection converges: for any  $u^0 \in K$ , the sequence  $(u^n)$  defined by (10.106) converges to the solution  $u$  of (10.102).*

**Proof.** The proof reuses that of theorem 10.5.4 by observing that the proof of (10.101) shows more generally that the mapping  $v \mapsto v - \mu J'(v)$  is strictly contracting when  $0 < \mu < 2\alpha/C^2$ , that is to say that

$$\exists \gamma \in ]0, 1[, \quad \|(v - \mu J'(v)) - (w - \mu J'(w))\| \leq \gamma \|v - w\|.$$

Since the projection  $P_K$  is weakly contracting from (12.2), the mapping  $v \mapsto P_K(v - \mu J'(v))$  is strictly contracting, which proves the convergence of the sequence  $(u^n)$  defined by (10.106) to the solution  $u$  of (10.102).  $\square$

**Exercise 10.5.2** Take  $V = \mathbb{R}^N$  and  $K = \{x \in \mathbb{R}^N \text{ such that } \sum_{i=1}^N x_i = 1\}$ . Make explicit the orthogonal projection operator  $P_K$  and interpret formula (10.106) in this case in terms of Lagrange multipliers.

### Uzawa's Algorithm

The preceding result shows that the gradient algorithm with fixed step and projection is applicable to a large class of convex optimization problems with constraints. But this conclusion is largely deluding from the practical point of view, since the projection operator  $P_K$  is not explicitly known in general: the projection of an element  $v \in V$  over an arbitrary convex closed set of  $V$  may be very difficult to determine!

An important exception concerns, in finite dimensions (for  $V = \mathbb{R}^M$ ), the subsets  $K$  of the form

$$K = \prod_{i=1}^M [a_i, b_i] \quad (10.107)$$

(with possibly  $a_i = -\infty$  or  $b_i = +\infty$  for certain indices  $i$ ). In effect, it is then easy to see that, if  $x = (x_1, x_2, \dots, x_M) \in \mathbb{R}^M$ ,  $y = P_K(x)$  has components

$$y_i = \min(\max(a_i, x_i), b_i) \quad \text{for } 1 \leq i \leq M, \quad (10.108)$$

in other words, it is enough just to ‘truncate’ the components of  $x$ . This simple property, together with the remarks about the duality stated in section 10.3, will lead us to a new algorithm. In effect, even if the primal problem involves a set  $K$  of admissible solutions over which the projection  $P_K$  cannot be explicitly given, the dual problem will frequently be posed over a set of the form (10.107), typically over  $(\mathbb{R}_+)^M$ . In this case, the dual problem may be solved by the gradient method with fixed step and projection, and the solution of the primal problem can then be obtained by solving a minimization problem **without constraint**. These remarks are the basis of Uzawa’s algorithm, which is in fact a method of looking for a saddle point.

Let us consider the convex minimization problem

$$\inf_{F(v) \leq 0} J(v), \quad (10.109)$$

where  $J$  is a convex functional defined over  $V$  and  $F$  a convex function of  $V$  over  $\mathbb{R}^M$ . Under the hypotheses of the Kuhn–Tucker theorem 10.3.4, the solution of (10.109) reduces to finding a saddle point  $(u, p)$  of the Lagrangian

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v), \quad (10.110)$$

over  $V \times (\mathbb{R}_+)^M$ . Starting from the definition 10.3.1 of the saddle point

$$\forall q \in (\mathbb{R}_+)^M \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall v \in V, \quad (10.111)$$

we deduce that  $(p - q) \cdot F(u) \geq 0$  for all  $q \in (\mathbb{R}_+)^M$ , from which we take, for all real  $\mu > 0$ ,

$$(p - q) \cdot (p - (p + \mu F(u))) \leq 0 \quad \forall q \in (\mathbb{R}_+)^M,$$

which, from (12.1), shows that

$$p = P_{\mathbb{R}_+^M}(p + \mu F(u)) \quad \forall \mu > 0, \quad (10.112)$$

$P_{\mathbb{R}_+^M}$  denotes the projection of  $\mathbb{R}^M$  over  $(\mathbb{R}_+)^M$ .

In view of this property and of the second inequality in (10.111), we can introduce **Uzawa’s algorithm**: starting from an arbitrary element  $p^0 \in (\mathbb{R}_+)^M$ , we construct the sequences  $(u^n)$  and  $(p^n)$  determined by the iterations

$$\begin{aligned} \mathcal{L}(u^n, p^n) &= \inf_{v \in V} \mathcal{L}(v, p^n), \\ p^{n+1} &= P_{\mathbb{R}_+^M}(p^n + \mu F(u^n)), \end{aligned} \quad (10.113)$$

$\mu$  being a fixed positive parameter.

**Theorem 10.5.9** *We assume that  $J$  is  $\alpha$ -convex and differentiable, that  $F$  is convex and Lipschitzian from  $V$  into  $\mathbb{R}^M$ , that is to say that there exists a constant  $C$  such that*

$$\|F(v) - F(w)\| \leq C\|v - w\| \quad \forall v, w \in V, \quad (10.114)$$

*and that there exists a saddle point  $(u, p)$  of the Lagrangian (10.110) over  $V \times (\mathbb{R}_+)^M$ . Then, if  $0 < \mu < 2\alpha/C^2$ , Uzawa's algorithm converges: for any initial element  $p^0$ , the sequence  $(u^n)$  defined by (10.113) converges to the solution  $u$  of problem (10.109).*

**Proof.** Let us recall first that the existence of a solution  $u$  of (10.109) follows from that of the saddle point  $(u, p)$  (see proposition 10.3.2), then that its uniqueness is a consequence of the  $\alpha$ -convexity of  $J$ . Likewise,  $p^n$  being fixed, the minimization problem in (10.113) has a unique solution  $u^n$ . From exercise 10.2.6, the Euler inequalities satisfied by  $u$  and  $u^n$  are written

$$\langle J'(u), v - u \rangle + p \cdot (F(v) - F(u)) \geq 0 \quad \forall v \in V, \quad (10.115)$$

$$\langle J'(u^n), v - u^n \rangle + p^n \cdot (F(v) - F(u^n)) \geq 0 \quad \forall v \in V. \quad (10.116)$$

Taking successively  $v = u^n$  in (10.115) and  $v = u$  in (10.116) and adding, we obtain

$$\langle J'(u) - J'(u^n), u^n - u \rangle + (p - p^n) \cdot (F(u^n) - F(u)) \geq 0,$$

from which by using the  $\alpha$ -convexity of  $J$  and by setting  $r^n = p^n - p$

$$r^n \cdot (F(u^n) - F(u)) \leq -\alpha\|u^n - u\|^2. \quad (10.117)$$

On the other hand, the projection  $P_{\mathbb{R}_+^M}$  being weakly contracting from (12.2), by subtracting (10.112) from (10.113) we obtain

$$\|r^{n+1}\| \leq \|r^n + \mu(F(u^n) - F(u))\|,$$

or

$$\|r^{n+1}\|^2 \leq \|r^n\|^2 + 2\mu r^n \cdot (F(u^n) - F(u)) + \mu^2 \|F(u^n) - F(u)\|^2.$$

Using (10.114) and (10.117), this becomes

$$\|r^{n+1}\|^2 \leq \|r^n\|^2 + (C^2\mu^2 - 2\mu\alpha)\|u^n - u\|^2.$$

If  $0 < \mu < 2\alpha/C^2$ , we can find  $\beta > 0$  such that  $C^2\mu^2 - 2\mu\alpha < -\beta$ , from which

$$\beta\|u^n - u\|^2 \leq \|r^n\|^2 - \|r^{n+1}\|^2. \quad (10.118)$$

This shows that the sequence  $\|r^n\|^2$  is decreasing: the right-hand side of (10.118) therefore tends to 0, which implies that  $u^n$  tends to  $u$ .  $\square$

Thus, Uzawa's algorithm allows us to approximate the solution of (10.109) by replacing this problem with constraints by a sequence of minimization problems without

constraints (10.113). At each iteration, the calculation of  $p^n$  is elementary, since we have simply

$$p_i^{n+1} = \max(p_i^n + \mu F_i(u^n), 0) \quad \text{for } 1 \leq i \leq M,$$

from (10.108). We must also note that theorem 10.5.9 says nothing about the convergence of the sequence  $(p^n)$ . In fact, this convergence is not ensured under the hypotheses of the theorem, which also does not ensure the uniqueness of the element  $p \in (\mathbb{R}_+)^M$  such that  $(u, p)$  is a saddle point (see remark 10.3.12 and exercise 10.5.3).

We have to make the link between Uzawa's algorithm and duality theory, as we have already stated. Let us recall first that the dual problem of (10.109) is written

$$\sup_{q \geq 0} \mathcal{G}(q), \quad (10.119)$$

where, by definition

$$\mathcal{G}(q) = \inf_{v \in V} \mathcal{L}(v, q), \quad (10.120)$$

and that the Lagrange multiplier  $p$  is a solution of the dual problem (10.119). In fact, under quite general hypotheses, we can show that  $\mathcal{G}$  is differentiable and that the gradient  $\mathcal{G}'(q)$  is precisely equal to  $F(u_q)$ , where  $u_q$  is the unique solution of the minimization problem (10.120). In effect, formally we have

$$\mathcal{G}(q) = J(u_q) + q \cdot F(u_q),$$

and differentiating with respect to  $q$

$$\mathcal{G}'(q) = F(u_q) + (J'(u_q) + q \cdot F'(u_q)) u'_q = F(u_q),$$

because of the optimality condition for  $u_q$ . We then see that **Uzawa's algorithm is none other than the gradient algorithm with fixed step and projection applied to the dual problem** since the second equation of (10.113) can be written  $p^{n+1} = P_{\mathbb{R}_+^M}(p^n + \mu \mathcal{G}'(p^n))$  (the change of sign with respect to (10.106) comes from the fact that the dual problem (10.119) is a problem of maximization and not of minimization). The reader will verify this assertion very easily in the particular case studied in the following exercise.

**Exercise 10.5.3** Apply Uzawa's algorithm to the problem of remark 10.3.12 (quadratic functional and affine constraints in finite dimensions). If the matrix  $B$  has range  $M$ , which assures the uniqueness of  $p$  from remark 10.3.12, show that the sequence  $p^n$  converges to  $p$ .

### Penalization of constraints

We conclude this subsection by briefly describing another way to approximate a minimization problem with constraints by a sequence of minimization problems

without constraints; this is the procedure of **penalization** of constraints. We avoid talking here of a ‘method’ or an ‘algorithm’ since penalization of the constraints is not, properly speaking, a method. The solution of the problems without constraints that we shall construct must be done with the help of the algorithms of Section 10.5.2. This solution can raise some difficulties, since the ‘penalized’ problem (10.122) is often ‘ill conditioned’ (see Section 13.1.2).

For simplicity we shall take the case where  $V = \mathbb{R}^N$ , and we again consider the convex minimization problem

$$\inf_{F(v) \leq 0} J(v), \quad (10.121)$$

where  $J$  is a continuous convex function from  $\mathbb{R}^N$  into  $\mathbb{R}$  and  $F$  a continuous convex function from  $\mathbb{R}^N$  into  $\mathbb{R}^M$ .

For  $\varepsilon > 0$ , we then introduce the problem without constraints

$$\inf_{v \in \mathbb{R}^N} \left( J(v) + \frac{1}{\varepsilon} \sum_{i=1}^M [\max(F_i(v), 0)]^2 \right), \quad (10.122)$$

where the constraints  $F_i(v) \leq 0$  are ‘penalized’. We can then state the following result, which shows that, for  $\varepsilon$  small, the problem (10.122) ‘approximates well’ the problem (10.121).

**Proposition 10.5.10** *We assume that  $J$  is continuous, strictly convex, and is infinite at infinity, that the functions  $F_i$  are convex and continuous for  $1 \leq i \leq M$ , and that the set*

$$K = \{v \in \mathbb{R}^N, \quad F_i(v) \leq 0 \quad \forall i \in \{1, \dots, M\}\}$$

*is nonempty. Denoting by  $u$  the unique solution of (10.121) and, for  $\varepsilon > 0$ ,  $u_\varepsilon$  the unique solution of (10.122), we then have*

$$\lim_{\varepsilon \rightarrow 0} u_\varepsilon = u.$$

**Proof.** As the set  $K$  is convex and closed, the existence and uniqueness of  $u$  follows from theorem 9.1.3 and from the strict convexity of  $J$ . Further, the function  $G(v) = \sum_{i=1}^M [\max(F_i(v), 0)]^2$  is continuous and convex since the function from  $\mathbb{R}$  into  $\mathbb{R}$  which to each  $x$  associates  $\max(x, 0)^2$  is convex and increasing. We deduce that the functional  $J_\varepsilon(v) = J(v) + \varepsilon^{-1}G(v)$  is strictly convex, continuous, and is infinite at infinity since  $G(v) \geq 0$ , which implies the existence and uniqueness of  $u_\varepsilon$ . As  $G(u) = 0$ , we can write

$$J_\varepsilon(u_\varepsilon) = J(u_\varepsilon) + \frac{G(u_\varepsilon)}{\varepsilon} \leq J_\varepsilon(u) = J(u). \quad (10.123)$$

This shows that

$$J(u_\varepsilon) \leq J_\varepsilon(u_\varepsilon) \leq J(u), \quad (10.124)$$

and therefore that  $u_\varepsilon$  is bounded from the ‘infinite at infinity’ condition. We can therefore extract from the family  $(u_\varepsilon)$  a sequence  $(u_{\varepsilon_k})$  which converges to a limit  $u_*$  when  $\varepsilon_k$  tends to 0. We then have  $0 \leq G(u_{\varepsilon_k}) \leq \varepsilon_k(J(u) - J(u_{\varepsilon_k}))$  from (10.123). Passing to the limit, we obtain  $G(u_*) = 0$ , which shows that  $u_* \in K$ . As (10.124) implies that  $J(u_*) \leq J(u)$ , we then have  $u_* = u$ , which concludes the proof, all the extracted sequences  $(u_{\varepsilon_k})$  converge to the same limit  $u$ .  $\square$

**Exercise 10.5.4** In addition to the hypotheses of proposition 10.5.10, we assume that the functions  $J$  and  $F_1, \dots, F_M$  are continuously differentiable. We denote by  $I(u)$  the set of active constraints at  $u$ , and we assume that the constraints are qualified at  $u$  in the sense of definition 10.2.13. Finally, we assume that the vectors  $(F'_i(u))_{i \in I(u)}$  are linearly independent, which ensures the uniqueness of the Lagrange multipliers  $\lambda_1, \dots, \lambda_M$  such that  $J'(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0$ , with  $\lambda_i = 0$  if  $i \notin I(u)$ . Show then that, for each index  $i \in \{1, \dots, M\}$

$$\lim_{\varepsilon \rightarrow 0} \left[ \frac{2}{\varepsilon} \max(F_i(u_\varepsilon), 0) \right] = \lambda_i.$$

**Remark 10.5.11** We see in Section 11.2.3 another method of penalization by the introduction of ‘barrier’ functions.  $\bullet$

## 10.5.4 Newton’s method

We work in finite dimensions  $V = \mathbb{R}^N$ . We shall explain the principle of Newton’s method to find the zeroes or roots of a function  $F$  of class  $C^2$  from  $\mathbb{R}^N$  into  $\mathbb{R}^N$ . Let  $u$  be a simple zero of  $F$  that is to say that

$$F(u) = 0 \quad \text{and} \quad F'(u) \text{ an invertible matrix.}$$

A Taylor formula in the neighbourhood of  $v$  gives us

$$F(u) = F(v) + F'(v)(u - v) + \mathcal{O}(\|u - v\|^2),$$

that is to say

$$u = v - (F'(v))^{-1} F(v) + \mathcal{O}(\|v - u\|^2).$$

Newton’s method consists of iteratively solving this equation neglecting the remainder. For an initial choice  $u^0 \in \mathbb{R}^N$ , we calculate

$$u^{n+1} = u^n - (F'(u^n))^{-1} F(u^n) \quad \text{for } n \geq 0. \quad (10.125)$$

Let us recall that we do not calculate the inverse of the matrix  $F'(u^n)$  in (10.125) but that we solve a linear system by one of the methods described in Section 13.1. From the point of view of optimization, Newton’s method is interpreted in the following way. Let  $J$  be a function of class  $C^3$  from  $\mathbb{R}^N$  into  $\mathbb{R}$ , and let  $u$  be a local minimum of  $J$ . If we set  $F = J'$ , we can apply the preceding method to solve the necessary

optimality condition  $J'(u) = 0$ . However, we can also see Newton's method as a minimization method. Because of the Taylor expansion

$$J(w) = J(v) + J'(v) \cdot (w - v) + \frac{1}{2} J''(v)(w - v) \cdot (w - v) + \mathcal{O}(\|w - v\|^3), \quad (10.126)$$

we can approximate  $J(w)$  in the neighbourhood of  $v$  by a quadratic function. Newton's method consists then of minimizing this approximation and iterating. The minimum of the quadratic part of the right-hand side of (10.126) is given by  $w = v - (J''(v))^{-1} J'(v)$  if the matrix  $J''(v)$  is positive definite. We then recover the iterative formula (10.125).

The principal advantage of Newton's method is its convergence which is more rapid than the preceding methods.

**Proposition 10.5.12** *Let  $F$  be a function of class  $C^2$  from  $\mathbb{R}^N$  into  $\mathbb{R}^N$ , and  $u$  a simple zero of  $F$  (that is,  $F(u) = 0$  and  $F'(u)$  is invertible). There exists a real  $\epsilon > 0$  such that, if  $u^0$  is close to  $u$  in the sense where  $\|u - u^0\| \leq \epsilon$ , Newton's method defined by (10.125) converges, that is to say that the sequence  $(u^n)$  converges to  $u$ , and there exists a constant  $C > 0$  such that*

$$\|u^{n+1} - u\| \leq C\|u^n - u\|^2. \quad (10.127)$$

**Proof.** By continuity of  $F'$  there exists  $\epsilon > 0$  such that  $F'$  is invertible at every point of the ball with centre  $u$  and radius  $\epsilon$ . Let us assume that  $u^n$  lies near to  $u$ , in the sense that  $\|u - u^n\| \leq \epsilon$ , therefore  $F'(u^n)$  is invertible. As  $F(u) = 0$ , we deduce from (10.125)

$$u^{n+1} - u = u^n - u - (F'(u^n))^{-1} (F(u^n) - F(u))$$

which, by Taylor expansion around  $u^n$ , becomes

$$u^{n+1} - u = (F'(u^n))^{-1} \mathcal{O}(\|u^n - u\|^2).$$

As  $\|u - u^n\| \leq \epsilon$ , we deduce that there exists a constant  $C > 0$  (independent of  $n$  and linked to the modulus of continuity of  $F'$  and  $F''$  over the ball of centre  $u$  and radius  $\epsilon$ ) such that

$$\|u^{n+1} - u\| \leq C\|u^n - u\|^2. \quad (10.128)$$

If  $\epsilon$  is sufficiently small so that  $C\epsilon \leq 1$ , we deduce from (10.128) that  $u^{n+1}$  remains in the ball of centre  $u$  and radius  $\epsilon$ . This allows us to verify, by recurrence, the hypothesis that  $\|u - u^n\| \leq \epsilon$  for all  $n \geq 0$ , and we have the conclusion.  $\square$

**Remark 10.5.13** Of course, we must remember that each iteration of Newton's method (10.125) needs the solution of a linear system, which is costly. Further, the rapid ('quadratic') convergence given by (10.127) only holds if  $F$  is of class  $C^2$ , and if  $u^0$  is quite close to  $u$ , which are more restrictive hypotheses than those we have used up until now. Effectively, even in a very simple case in  $\mathbb{R}$ , Newton's method



can diverge for certain initial data  $u^0$ ; we must also note that the quadratic convergence (10.127) only happens in the neighbourhood of a simple zero, as is seen in the application of Newton's method to the function  $F(x) = \|x\|^2$  in  $\mathbb{R}^N$ , for which the convergence is only geometrical. In addition, if we apply Newton's method to the minimization of a function  $J$  as explained above, it may be that the method converges to a maximum or a saddle point of  $J$ , and does not tend to a minimum, since it only looks for the zeros of  $J'$ . Newton's method is therefore not better in all ways to the preceding algorithms, but the property of local quadratic convergence (10.127), however, makes it particularly interesting. •

**Remark 10.5.14** A major drawback of Newton's method is the need to know the Hessian  $J''(v)$  (or the derivative matrix  $F'(v)$ ). When the problem is very large or if  $J$  is not easily twice differentiable, we can modify Newton's method to avoid the calculation of this matrix  $J''(v) = F'(v)$ . The methods, called quasi-Newton, propose iteratively calculating an approximation  $S^n$  of  $(F'(u^n))^{-1}$ . We replace the formula (10.125) by

$$u^{n+1} = u^n - S^n F(u^n) \quad \text{for } n \geq 0.$$

In general, we calculate  $S^n$  by a recurrence formula of the type

$$S^{n+1} = S^n + C^n$$

where  $C^n$  is a matrix of rank 1 which depends on  $u^n, u^{n+1}, F(u^n), F(u^{n+1})$ , chosen so that  $S^n - (F'(u^n))^{-1}$  converges to 0. For more details about these quasi-Newton methods we refer to [3] and [30]. •

We can adapt Newton's method to the minimization of a function  $J$  with equality constraints. Let  $J$  be a function of class  $C^3$  from  $\mathbb{R}^N$  into  $\mathbb{R}$ ,  $G = (G_1, \dots, G_M)$  a function of class  $C^3$  from  $\mathbb{R}^N$  into  $\mathbb{R}^M$  (with  $M \leq N$ ), and let  $u$  be a local minimum of

$$\min_{v \in \mathbb{R}^N, G(v)=0} J(v). \quad (10.129)$$

If the vectors  $(G'_1(u), \dots, G'_M(u))$  are linearly independent, the necessary optimality condition of theorem 10.2.8 is

$$J'(u) + \sum_{i=1}^M \lambda_i G'_i(u) = 0, \quad G_i(u) = 0 \quad 1 \leq i \leq M. \quad (10.130)$$

where the  $\lambda_1, \dots, \lambda_M \in \mathbb{R}$  are the Lagrange multipliers. We can then solve the system (10.130) of  $(N + M)$  equations with  $(N + M)$  unknowns  $(u, \lambda) \in \mathbb{R}^{N+M}$  by a Newton method. We therefore set

$$F(u, \lambda) = \begin{pmatrix} J'(u) + \lambda \cdot G'(u) \\ G(u) \end{pmatrix},$$

whose matrix derivative is

$$F'(u, \lambda) = \begin{pmatrix} J''(u) + \lambda \cdot G''(u) & (G'(u))^* \\ G'(u) & 0 \end{pmatrix}.$$

We can then apply the Newton algorithm (10.125) to this function  $F(u, \lambda)$  if the matrix  $F'(u, \lambda)$  is invertible. We shall see that this condition is 'natural' in the sense

that it corresponds to a slightly stronger version of the second order optimality condition of proposition 10.2.11. The matrix  $F'(u, \lambda)$  is invertible if it is injective. Let  $(w, \mu)$  be an element of its kernel

$$\begin{cases} J''(u)w + \lambda \cdot G''(u)w + (G'(u))^* \mu = 0 \\ G'_i(u) \cdot w = 0 \quad \text{for } 1 \leq i \leq M \end{cases}$$

We deduce that

$$w \in \text{Ker} G'(u) = \bigcap_{i=1}^M \text{Ker} G'_i(u) \quad \text{and} \quad (J''(u) + \lambda \cdot G''(u))w \in \text{Im}(G'(u))^*.$$

Now  $\text{Im}(G'(u))^* = [\text{Ker} G'(u)]^\perp$ . Consequently, if we assume that

$$(J''(u) + \lambda \cdot G''(u))(w, w) > 0 \quad \forall w \in \text{Ker} G'(u), \quad w \neq 0, \quad (10.131)$$

the matrix  $F'(u, \lambda)$  is invertible. We remark that (10.131) is the strict inequality in the second order optimality condition of proposition 10.2.11. It is therefore natural to make the hypothesis (10.131) which allows us to use the Newton algorithm. We can therefore prove the convergence of this method (see [3]).

It is interesting to interpret this algorithm as a minimization method. We introduce the Lagrangian  $\mathcal{L}(v, \mu) = J(v) + \mu \cdot G(v)$ , its derivatives with respect to  $v$ ,  $\mathcal{L}'$ , and  $\mathcal{L}''$ , and we verify that the equation

$$(u^{n+1}, \lambda^{n+1}) = (u^n, \lambda^n) - (F'(u^n, \lambda^n))^{-1} F(u^n, \lambda^n)$$

is the optimality condition so that  $u^{n+1}$  is a minimum point of the quadratic problem with affine constraints

$$\min_{\substack{w \in \mathbb{R}^N \\ G(u^n) + G'(u^n) \cdot (w - u^n) = 0}} Q^n(w), \quad (10.132)$$

with

$$Q^n(w) = \left( \mathcal{L}(u^n, \lambda^n) + \mathcal{L}'(u^n, \lambda^n) \cdot (w - u^n) + \frac{1}{2} \mathcal{L}''(u^n, \lambda^n)(w - u^n) \cdot (w - u^n) \right),$$

and  $\lambda^{n+1}$  is the Lagrange multiplier associated with the minimum point of (10.132). We remark that in (10.132) we have made a Taylor expansion of order two at  $w$  of the Lagrangian  $\mathcal{L}(w, \lambda^n)$  and we have linearized the constraint  $G(w)$  around the point  $u^n$ .

**Remark 10.5.15** In (10.132) we have used a quadratic approximation of the Lagrangian and not of the function  $J$ . We could try the following iterative method for the solution of the quadratic approximation with affine constraints

$$\min_{\substack{w \in \mathbb{R}^N \\ G(v) + G'(v) \cdot (w - v) = 0}} \left( J(v) + J'(v) \cdot (w - v) + \frac{1}{2} J''(v)(w - v) \cdot (w - v) \right). \quad (10.133)$$

Unfortunately, the method based on (10.133) cannot converge! In particular, it is not obvious that the Hessian  $J''(v)$  is positive definite over the space of constraints (it is the Hessian of the Lagrangian which is positive as is affirmed by the second order optimality condition of proposition 10.2.11). •

*This page intentionally left blank*

# 11 Methods of operational research

(Written in collaboration with Stéphane Gaubert)

---

## 11.1 Introduction

In this chapter, we present several tools from OR (operational research). In OR, the word ‘operational’ was initially used in its proper sense: OR was born, mainly, from planning problems which arose during the Second World War and shortly after. Thus G. Dantzig, the inventor of the simplex algorithm, was an adviser for the American airforce, and the planning of the Berlin airlift in 1948 is a famous application of linear programming (see [31] for more detail). Since then the domain has been considerably developed and civilized. The problems of OR abound in industry and the service sector: we can cite, for example, timetabling problems (for aircrews, for the employees of a call centre, etc.), routing of vehicles, networks, siting of warehouses, stock management, workshop scheduling, etc. OR borrows tools from many scientific fields: continuous optimization and combinatorial optimization, but also from discrete mathematics and in particular from graph theory; from computing, on the one hand via complexity theory, which allows us to identify ‘easy’ problems, which are solvable in polynomial time in the size of the problem, from difficult problems, and on the other hand via constraint programming, or ‘CP’ (which allows us to enumerate the solutions intelligently). Some questions in OR are also linked to probability theory (for example, to understand stochastic optimization algorithms such as simulated annealing), to computing or game theory (for dynamical decision problems). A significant part of the activity in OR is linked to the practice (modelling, heuristics, etc.). Our intention in this chapter is not to present OR, but rather to consider a mathematical part of the field, combinatorial optimization, which has strong ties to the continuous optimization treated in the preceding chapters: the most efficient methods for the exact solution of combinatorial problems are often based on convex programming, and in return,

considering discrete problems and objects (problems of flows, assignment problems extreme points of polyhedra, electrical networks, Laplacians of graphs, etc.), often allow better understanding of the analogues of these problems and objects which occur in analysis.

In this chapter we shall present five important methods. **Linear programming**, which will be the subject of Section 11.2, will allow us to solve efficiently continuous optimization problems where the constraints and the criterion are expressed linearly, which is very common in OR. Sometimes, it is essential to find an integer solution. We shall see in Section 11.3, with the help of the idea of an **integer polyhedron**, that this can be done without increasing the complexity in certain special cases, which includes the important class of flow problems. Section 11.4 presents another method, **dynamic programming**, which is naturally adapted to problems of shortest path, and which is also useful in more difficult combinatorial problems. Section 11.5 gives an example of a **greedy algorithm**: greedy algorithms are only optimal for very particular problems, but they are often useful to produce heuristics. We shall see finally in Section 11.6, that when the preceding tools cannot be applied directly, it is often possible to obtain an optimal solution by **separation and evaluation** ('branch and bound'), that is, by a tree search combined with an approximation of the problem.

## 11.2 Linear programming

We have not yet said anything about example 9.1.1 which is typical of a very large class of problems, called linear programming problems. Because of the practical importance of these problems we shall now devote a whole section to their study.

### 11.2.1 Definitions and properties

We want to solve the following **linear programming problem**,

$$\inf_{x \in \mathbb{R}^n \text{ such that } Ax=b, x \geq 0} c \cdot x, \quad (11.1)$$

where  $A$  is a matrix of size  $m \times n$ ,  $b \in \mathbb{R}^m$ ,  $c \in \mathbb{R}^n$ , and the constraint  $x \geq 0$  means that all the components of  $x$  are positive or zero. In everything that follows we shall assume that  $m \leq n$  and that the range of  $A$  is exactly  $m$ . In effect, if  $\text{rg}(A) < m$ , certain rows of  $A$  are linked and two possibilities arise: either the constraints (corresponding to these rows) are incompatible, or they are redundant and we can therefore eliminate the extra rows.

Problem (11.1) seems to be a particular case of linear programming since the inequality constraints are only of the type  $x \geq 0$ . This is not the case, and every linear programming problem of the type

$$\inf_{x \in \mathbb{R}^n \text{ such that } Ax \geq b, A'x=b'} c \cdot x.$$

can be put into the standard form (11.1) by rescaling the data. In effect, let us remark first of all that the equality constraints  $A'x = b'$  are obviously equivalent to the inequality constraints  $A'x \leq b'$  and  $A'x \geq b'$ . We can therefore restrict ourselves to the following case (which only contains inequality constraints)

$$\inf_{x \in \mathbb{R}^n \text{ such that } Ax \geq b} c \cdot x. \quad (11.2)$$

In (11.2) we can replace the inequality constraint by introducing new variables, called **slack** variables,  $\lambda \in \mathbb{R}^m$ . The inequality constraint  $Ax \geq b$  is then equivalent to  $Ax = b + \lambda$  with  $\lambda \geq 0$ . Thus (11.2) is equivalent to

$$\inf_{(x, \lambda) \in \mathbb{R}^{(n+m)} \text{ such that } Ax = b + \lambda, \lambda \geq 0} c \cdot x. \quad (11.3)$$

Finally, if we decompose each component of  $x$  into its positive and negative parts, that is if we set  $x = x^+ - x^-$  with  $x^+ = \max(0, x)$  and  $x^- = -\min(0, x)$ , we obtain that (11.2) is equivalent to

$$\inf_{(x^+, x^-, \lambda) \in \mathbb{R}^{(2n+m)} \text{ such that } Ax^+ - Ax^- = b + \lambda, x^+ \geq 0, x^- \geq 0, \lambda \geq 0} c \cdot (x^+ - x^-). \quad (11.4)$$

which is in the standard form (but with more variables). There is therefore no loss of generality in studying the standard linear programming problem (11.1).

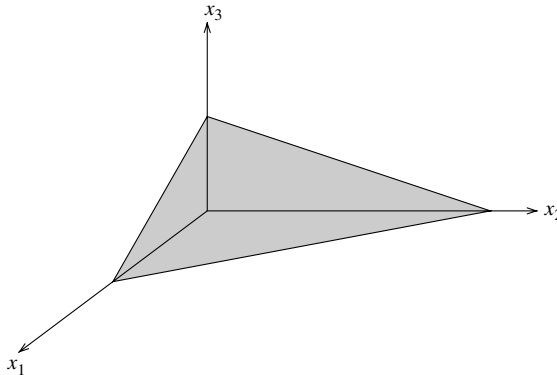


Figure 11.1. Admissible set for example (11.5).

We have already given a concrete motivation to linear programming at the beginning of Chapter 9 (see example 9.1.1). Let us consider for the moment a

simple example which will allow us to understand some essential aspects of linear programming

$$\min_{\substack{x_1 \geq 0, x_2 \geq 0, x_3 \geq 0 \\ 2x_1 + x_2 + 3x_3 = 6}} x_1 + 4x_2 + 2x_3. \quad (11.5)$$

On Figure 11.1 we have traced the set of  $(x_1, x_2, x_3)$  which satisfies the constraints: a plane triangle  $T$ . This is a closed compact set of  $\mathbb{R}^3$ , therefore the continuous function  $x_1 + 4x_2 + 2x_3$  attains its minimum that we denote  $M$ . To determine this minimum we can consider the family of parallel planes  $x_1 + 4x_2 + 2x_3 = c$  parameterized by  $c$ . By increasing the value of  $c$  starting from  $-\infty$ , we 'sweep' the space  $\mathbb{R}^3$  until we reach the triangle  $T$ , and the minimum  $M$  is attained when the plane 'touches' this triangle. In other words, every minimum point of (11.5) is on the boundary of the triangle  $T$ . Another way to see this is to say that the function  $x_1 + 4x_2 + 2x_3$  has a nonzero gradient in  $T$  therefore its extrema are found on the boundary of  $T$ . For example (11.5), the (unique) minimum point is the vertex  $(0, 3, 0)$  of  $T$ . We shall see that this is a general fact: a minimum point (if it exists) can always be found at one of the vertices of the geometrical set of vectors  $x$  which satisfy the constraints. It is 'enough' then to enumerate all the vertices in order to find the minimum: it is exactly this that is (in an intelligent way) the simplex algorithm that we see in the next section.

To establish this property generally for the standard linear programming problem (11.1), we need some definitions which allow us to specify the vocabulary.

**Definition 11.2.1** *The set  $X_{\text{ad}}$  of vectors of  $\mathbb{R}^n$  which satisfy the constraints of (11.1), that is*

$$X_{\text{ad}} = \{x \in \mathbb{R}^n \text{ such that } Ax = b, x \geq 0\},$$

*is called the set of **admissible solutions**. We say vertex or extremal point of  $X_{\text{ad}}$  for every point  $\bar{x} \in X_{\text{ad}}$  which cannot be decomposed into a (nontrivial) convex combination of two other points of  $X_{\text{ad}}$ , that is, if there exist  $y, z \in X_{\text{ad}}$  and  $\theta \in ]0, 1[$  such that  $\bar{x} = \theta y + (1 - \theta)z$ , then  $y = z = \bar{x}$ .*

**Remark 11.2.2** The vocabulary of optimization is misleading for neophytes. We say (admissible) solution for a vector which satisfies the constraints. Conversely, a vector which attains the minimum of (11.1) is called the **optimal solution** (or minimum point). •

We easily verify that the set  $X_{\text{ad}}$  is a **polyhedron** (possibly empty). (Let us recall that a polyhedron is a finite intersection of halfspaces of  $\mathbb{R}^n$ .) Its extreme points are therefore the vertices of this polyhedron. When  $X_{\text{ad}}$  is empty, by convention we note that

$$\inf_{x \in \mathbb{R}^n \text{ such that } Ax=b, x \geq 0} c \cdot x = +\infty.$$

**Lemma 11.2.3** *There exists at least one optimal solution (or minimum point) of the standard linear programming problem (11.1) if and only if the minimum value is finite*

$$-\infty < \inf_{x \in \mathbb{R}^n \text{ such that } Ax=b, x \geq 0} c \cdot x < +\infty.$$

**Proof.** Let  $(x^k)_{k \geq 1}$  be a minimizing sequence of (11.1). We introduce the matrix  $\mathcal{A}$  defined by

$$\mathcal{A} = \begin{pmatrix} c^* \\ A \end{pmatrix}.$$

The sequence  $\mathcal{A}x^k$  belongs to the following cone

$$C = \left\{ \sum_{i=1}^n x_i \mathcal{A}_i \quad \text{with } x_i \geq 0 \right\},$$

where the  $\mathcal{A}_i$  are the columns of the matrix  $\mathcal{A}$ . From the Farkas lemma 10.2.17 the cone  $C$  is closed, which implies that

$$\lim_{k \rightarrow +\infty} \mathcal{A}x^k = \begin{pmatrix} z_0 \\ b \end{pmatrix} \in C,$$

therefore there exists  $\bar{x} \geq 0$  such that

$$\begin{pmatrix} z_0 \\ b \end{pmatrix} = \begin{pmatrix} c \cdot \bar{x} \\ A\bar{x} \end{pmatrix},$$

and the minimum is attained at  $\bar{x}$ . □

**Definition 11.2.4** *We say **basis** associated with (11.1) for a basis of  $\mathbb{R}^m$  composed of  $m$  columns of  $A$ . We denote this basis by  $B$  which is a submatrix of  $A$ , which is square of order  $m$  and invertible. After permutation of its columns we can write  $A$  in the form  $(B, N)$  where  $N$  is a matrix of size  $m \times (n - m)$ . In the same way we can decompose  $x$  into  $(x_B, x_N)$  so that we have*

$$Ax = Bx_B + Nx_N.$$

*The components of the vector  $x_B$  are called basic variables and those of  $x_N$  nonbasic variables. A **basic solution** is a vector  $x \in X_{\text{ad}}$  such that  $x_N = 0$ . If moreover one of the components of  $x_B$  is zero, we say that the basic solution is **degenerate**.*

The idea of a basic solution corresponds to that of a vertex of  $X_{\text{ad}}$ .

**Lemma 11.2.5** *The vertices of the polyhedron  $X_{\text{ad}}$  are exactly the basic solutions.*



**Proof.** If  $x \in X_{\text{ad}}$  is a basic solution, in a certain basis of  $\mathbb{R}^n$  we have  $x = (x_1, \dots, x_m, 0, \dots, 0)$  and there exists a basis  $(b_1, \dots, b_m)$  of  $\mathbb{R}^m$  such that  $\sum_{i=1}^m x_i b_i = b$ . Let us assume that there exists  $0 < \theta < 1$  and  $y, z \in X_{\text{ad}}$  such that  $x = \theta y + (1 - \theta)z$ . Necessarily, the  $n - m$  last components of  $y$  and  $z$  are zero and, as  $y$  and  $z$  belong to  $X_{\text{ad}}$ , we have  $\sum_{i=1}^m y_i b_i = b$  and  $\sum_{i=1}^m z_i b_i = b$ . By uniqueness of the decomposition in a basis, we deduce that  $x = y = z$ , and therefore  $x$  is a vertex of  $X_{\text{ad}}$ .

Conversely, if  $x$  is a vertex of  $X_{\text{ad}}$ , we denote by  $k$  the number of its nonzero components, and after a possible rearrangement we have  $b = \sum_{i=1}^k x_i a_i$  where the  $(a_i)$  are the columns of  $A$ . To show that  $x$  is a basic solution it is enough to prove that the family  $(a_1, \dots, a_k)$  is linearly independent in  $\mathbb{R}^m$  (we obtain a basis  $B$  by supplementing this family). Let us assume that this is not the case: there then exists  $y \neq 0$  such that  $\sum_{i=1}^k y_i a_i = 0$  and  $(y_{k+1}, \dots, y_n) = 0$ . As the components  $(x_1, \dots, x_k)$  are strictly positive, there exists  $\epsilon > 0$  such that  $(x + \epsilon y) \in X_{\text{ad}}$  and  $(x - \epsilon y) \in X_{\text{ad}}$ . The fact that  $x = (x + \epsilon y)/2 + (x - \epsilon y)/2$  contradicts the extremal character of  $x$ , therefore  $x$  is a basic solution.  $\square$

The following fundamental result tells us that it is sufficient to look for an optimal solution among the vertices of the polyhedron  $X_{\text{ad}}$ .

**Proposition 11.2.6** *If there exists an optimal solution of the standard linear programming problem (11.1), then there exists an optimal basic solution.*

**Proof.** The proof is very similar to that of lemma 11.2.5. Let  $x \in X_{\text{ad}}$  be an optimal solution of (11.1). We denote by  $k$  the number of its nonzero components, and after a possible rearrangement we have

$$b = \sum_{i=1}^k x_i a_i,$$

where the  $(a_i)$  are the columns of  $A$ . If the family  $(a_1, \dots, a_k)$  is linearly independent in  $\mathbb{R}^m$ , then  $x$  is an optimal basic solution. If  $(a_1, \dots, a_k)$  is linearly dependent, then there exists  $y \neq 0$  such that

$$\sum_{i=1}^k y_i a_i = 0 \quad \text{and} \quad (y_{k+1}, \dots, y_n) = 0.$$

As the components  $(x_1, \dots, x_k)$  are strictly positive, there exists  $\epsilon > 0$  such that  $(x \pm \epsilon y) \in X_{\text{ad}}$ . As  $x$  is a minimum point, we must have

$$c \cdot x \leq c \cdot (x \pm \epsilon y),$$

that is  $c \cdot y = 0$ . We then define a family of points  $z_\epsilon = x + \epsilon y$  parameterized by  $\epsilon$ . Starting from the value  $\epsilon = 0$ , if we increase or decrease  $\epsilon$  we stay in the set  $X_{\text{ad}}$  up to a value  $\epsilon_0$  beyond which the constraint  $z_\epsilon \geq 0$  is violated. In other words,  $z_{\epsilon_0} \in X_{\text{ad}}$  has more than  $(k - 1)$  nonzero components and is an optimal solution. We then repeat the preceding argument with  $x = z_{\epsilon_0}$  and a family of  $(k - 1)$  columns  $(a_i)$ . By decreasing the size of this family, we will finally obtain a linearly independent family and a basic optimal solution.  $\square$

**Remark 11.2.7** By applying proposition 11.2.6 when  $c = 0$  (every admissible solution is then optimal), we see thanks to lemma 11.2.5 that as soon as  $X_{\text{ad}}$  is nonempty,  $X_{\text{ad}}$  has at least one vertex. This property does not hold for general polyhedra (consider a half-plane of  $\mathbb{R}^2$ ). •

**Exercise 11.2.1** Solve the following linear programming problem

$$\max_{x_1 \geq 0, x_2 \geq 0} x_1 + 2x_2$$

under the constraints

$$\begin{cases} -3x_1 + 2x_2 & \leq 2 \\ -x_1 + 2x_2 & \leq 4 \\ x_1 + x_2 & \leq 5 \end{cases}$$

In practice, the number of vertices of the polyhedron  $X_{\text{ad}}$  is gigantic as they can be exponential with respect to the number of variables. We verify this in an example in the following exercise.

**Exercise 11.2.2** Show that we can choose the matrix  $A$  of size  $m \times n$  and the vector  $b \in \mathbb{R}^m$  in such a way that  $X_{\text{ad}}$  is the unit cube  $[0, 1]^{n-m}$  in the affine subspace of dimension  $n - m$  defined by  $Ax = b$ . Deduce that the number of vertices of  $X_{\text{ad}}$  is then  $2^{n-m}$ .

## 11.2.2 The simplex algorithm

The simplex algorithm was introduced by G. Dantzig in the 1940s. It consists of visiting the vertices of the polyhedron of admissible solutions until we find an optimal solution (which is guaranteed if the linear programming problem has an optimal solution). The simplex algorithm is not content with enumerating all the vertices, it decreases the value of the function  $c \cdot x$  passing from one vertex to the next.

We consider the standard linear programming problem (11.1). Let us recall that a vertex (or basic solution) of the set of admissible solutions  $X_{\text{ad}}$  is characterized by a basis  $B$  ( $m$  linearly independent columns of  $A$ ). After permutation of its columns, we can write

$$A = (B, N) \quad \text{and} \quad x = (x_B, x_N),$$

so that we have  $Ax = Bx_B + Nx_N$ . All admissible solutions can be written  $x_B = B^{-1}(b - Nx_N) \geq 0$  and  $x_N \geq 0$ . The vertex associated with  $B$  is defined (if it exists) by  $\bar{x}_N = 0$  and  $\bar{x}_B = B^{-1}b \geq 0$ . If we also decompose  $c = (c_B, c_N)$  in this basis, then we can compare the cost of an arbitrary admissible solution  $x$  with that of the basic solution  $\bar{x}$ . We easily see that  $c \cdot \bar{x} \leq c \cdot x$  if and only if

$$c_B \cdot B^{-1}b \leq c_B \cdot B^{-1}(b - Nx_N) + c_N \cdot x_N. \quad (11.6)$$

We deduce the following optimality condition.

**Proposition 11.2.8** *Let us assume that the basic solution associated with  $B$  is non-degenerate, that is,  $B^{-1}b > 0$ . A necessary and sufficient condition so that this basic solution associated with  $B$  is optimal is that*

$$\tilde{c}_N = c_N - N^*(B^{-1})^*c_B \geq 0. \quad (11.7)$$

The vector  $\tilde{c}_N$  is called the **vector of reduced costs**.

**Proof.** Let  $\bar{x}$  be a nondegenerate basic solution associated with  $B$ . If  $\tilde{c}_N \geq 0$ , then for every admissible solutions  $x$  we have

$$c \cdot x - c \cdot \bar{x} = \tilde{c}_N \cdot x_N \geq 0,$$

since  $x_N \geq 0$ . Thus the condition (11.7) is sufficient for  $\bar{x}$  to be optimal. Conversely, let us assume that there exists a component  $i$  of  $\tilde{c}_N$  which is strictly negative,  $(\tilde{c}_N \cdot e_i) < 0$ . For  $\epsilon > 0$  we then define a vector  $x(\epsilon)$  by  $x_N(\epsilon) = \epsilon e_i$  and  $x_B(\epsilon) = B^{-1}(b - Nx_N(\epsilon))$ . By construction  $Ax(\epsilon) = b$  and, as  $B^{-1}b > 0$ , for sufficiently small values of  $\epsilon$  we have  $x(\epsilon) \geq 0$ , therefore  $x(\epsilon) \in X_{\text{ad}}$ . On the other hand,  $x(0) = \bar{x}$  and, as  $\epsilon > 0$ , we have

$$c \cdot x(\epsilon) = c \cdot x(0) + \epsilon(\tilde{c}_N \cdot e_i) < c \cdot \bar{x},$$

which shows that  $\bar{x}$  is not optimal. Thus the condition (11.7) is necessary.  $\square$

**Remark 11.2.9** In the framework of proposition 11.2.8, if the basic solution is degenerate, condition (11.7) remains sufficient but is no longer necessary.  $\bullet$

We deduce from proposition 11.2.8 a practical method to decrease the value of the cost function  $c \cdot x$  starting from a basic solution  $\bar{x}$  (nondegenerate and nonoptimal). As  $\bar{x}$  is nonoptimal, there exists a component of the reduced cost vector  $\tilde{c}_N$  such that  $\tilde{c}_N \cdot e_i < 0$ . We then define  $x(\epsilon)$  as above. Since the cost decreases linearly with  $\epsilon$ , it is beneficial to take the largest possible value of  $\epsilon$  such that we remain in  $X_{\text{ad}}$ . This is the principle of the simplex algorithm that we now present.

### The simplex algorithm

- Initialization (phase I): We look for an initial basis  $B^0$  such that the associated basic solution  $x^0$  is admissible

$$x^0 = \begin{pmatrix} (B^0)^{-1}b \\ 0 \end{pmatrix} \geq 0.$$

- Iterations (phase II): at step  $k \geq 0$ , we have a basis  $B^k$  and an admissible basic solution  $x^k$ . We calculate the reduced cost  $\tilde{c}_N^k = c_N^k - (N^k)^*(B^k)^{-1}c_B^k$ . If  $\tilde{c}_N^k \geq 0$ , then  $x^k$  is optimal and the algorithm is finished. Otherwise, there exists a nonbasis variable with index  $i$  such that  $(\tilde{c}_N \cdot e_i) < 0$ , and we denote by  $a_i$  the corresponding column of  $A$ . We set

$$x^k(\epsilon) = (x_B^k(\epsilon), x_N^k(\epsilon)) \quad \text{with } x_N^k(\epsilon) = \epsilon e_i, \quad x_B^k(\epsilon) = (B^k)^{-1}(b - \epsilon a_i).$$

- Either we can choose  $\epsilon > 0$  as large as we want with  $x^k(\epsilon) \in X_{\text{ad}}$ . In this case, the minimum of the linear programming problem is  $-\infty$ .
- Or there exists a maximal value  $\epsilon^k$  and an index  $j$  such that the  $j$ th component of  $x^k(\epsilon^k)$  is zero. We thus obtain a new basis  $B^{k+1}$  deduced from  $B^k$  by replacing its  $j$ th column by the column  $a_j$ . The new admissible basic solution  $x^{k+1}$  has a cost less than or equal to that of  $x^k$ .

There are a certain number of practical points still to be specified in the simplex algorithm. We quickly review them.

### Degeneracy and cycling

We always have  $c \cdot x^{k+1} \leq c \cdot x^k$ , but we can have equality if the admissible basic solution  $x^k$  is degenerate, in which case we can find  $\epsilon^k = 0$  (if it is not degenerate, the proof of proposition 11.2.8 guarantees a strict inequality). We have therefore changed the basis without improving the cost: this is the phenomenon of cycling which can prevent the convergence of the algorithm. There are ways of stopping this, but in practice cycling never occurs.

In the absence of cycling, the simplex algorithm traverses a subset of vertices of  $X_{\text{ad}}$  (strictly) diminishing the cost. As there are a finite number of vertices, the algorithm must necessarily find an optimal vertex with minimal cost. We have therefore proved the following result.

**Lemma 11.2.10** *If all admissible basic solutions  $x^k$  produced by the simplex algorithm are nondegenerate, then the algorithm converges in a finite number of steps.*

*A priori* the number of iterations of the simplex algorithm can be as large as the number of vertices (which is exponential with respect to the number of variables  $n$ ; see exercise 11.2.2). Although there exist (academic) examples where this is effectively the case, in practice this algorithm converges in a number of steps which is a polynomial function of  $n$ .

### Choice of the change of basis

If there are many components of the reduced cost vector  $\tilde{c}_N^k$  which are strictly negative, we must make a choice in the algorithm. Many strategies are possible, but in general we choose the most negative.

### Initialization

How do we find a basic admissible solution at the initialization? (Let us recall that the condition of admissibility  $x_B = B^{-1}b \geq 0$  is not obvious in general.) We could know one because of the structure of the problem. For example, for the problem (11.4) which has  $m$  slack variables,  $-I_m$  is a basis of the ‘global’ matrix of the equality

constraints of (11.4). If  $b \leq 0$ , the vector  $(x^0, \lambda^0) = (0, -b)$  is then a basic admissible solution for (11.4).

In the general case, we introduce a new variable  $y \in \mathbb{R}^m$ , a new cost vector  $k > 0$ , and a new linear programming problem

$$\inf_{\substack{x \geq 0, \ y \geq 0 \\ Ax + y = b}} c \cdot x + k \cdot y \quad (11.8)$$

where we have already multiplied all the equality constraints corresponding to the negative components of  $b$  by  $-1$  so that  $b \geq 0$ . The vector  $(x^0, y^0) = (0, b)$  is a basic admissible solution for this problem. If there exists an admissible solution of the original linear programming problem (11.1) and if the minimum of (11.1) is not  $-\infty$ , then the optimal solutions of (11.8) must satisfy  $y = 0$ . By applying the simplex algorithm to (11.8), we find a basic admissible solution for (11.1) if one exists. If none exists (that is, if  $X_{\text{ad}} = \emptyset$ ), we detect this as the minimum of (11.8) is attained by a vector  $(x, y)$  with  $y \neq 0$ . Obviously, to quickly find an admissible solution for (11.1) it is in our interest to choose  $k$  very large with respect to  $c$ .

### Inversion of the basis

As we have described it, the simplex algorithm needs the inversion of  $B^k$  at each step, which can be very costly for large problems (with many constraints since the order of  $B^k$  is equal to the number of constraints). We can use the fact that  $B^{k+1}$  only differs from  $B^k$  by one column to develop a better strategy. In effect, if it is the  $j$ th column which changes, we have

$$B^{k+1} = B^k E^k \quad \text{with } E^k = \begin{pmatrix} 1 & & l_1 & & \\ & \ddots & \vdots & & 0 \\ & & 1 & & \\ & & & l_j & \\ & & & \vdots & 1 \\ 0 & & & \vdots & & \ddots \\ & & l_n & & & & 1 \end{pmatrix},$$

and  $E^k$  is easy to invert

$$(E^k)^{-1} = \frac{1}{l_j} \begin{pmatrix} 1 & & -l_1 & & \\ & \ddots & \vdots & & 0 \\ & & 1 & -l_{j-1} & \\ & & & 1 & \\ & & & -l_{j+1} & 1 \\ 0 & & & \vdots & & \ddots \\ & & -l_n & & & & 1 \end{pmatrix}.$$

We therefore use the formula, in a factorized form,

$$(B^k)^{-1} = (E^{k-1})^{-1}(E^{k-2})^{-1} \dots (E^0)^{-1}(B^0)^{-1}.$$

**Exercise 11.2.3** Solve, by the simplex algorithm, the linear programming problem

$$\min_{x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0, x_5 \geq 0} x_1 + 2x_2$$

under the constraints

$$\begin{cases} -3x_1 + 2x_2 + x_3 = 2 \\ -x_1 + 2x_2 + x_4 = 4 \\ x_1 + x_2 + x_5 = 5 \end{cases}$$

**Exercise 11.2.4** Solve, by the simplex algorithm, the linear programming problem

$$\min_{x_1 \geq 0, x_2 \geq 0} 2x_1 - x_2$$

under the constraints  $x_1 + x_2 \leq 1$  and  $x_2 - x_1 \leq 1/2$ .

**Exercise 11.2.5** Solve, by the simplex algorithm, the linear programming problem

$$\min_{x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0} 3x_3 - x_4$$

under the constraints

$$\begin{cases} x_1 - 3x_3 + 3x_4 = 6 \\ x_2 - 8x_3 + 4x_4 = 4 \end{cases}$$

### 11.2.3 Interior point algorithms

Since the work by Khachian and Karmarkar at the beginning of the 1980s, a new class of algorithms, called interior point algorithms, have emerged to solve linear programming problems. The name of this class of algorithms comes from the fact that it is the opposite of the simplex method (which, traversing the vertices, remains on the boundary of the polyhedron  $X_{\text{ad}}$ ). These interior point algorithms evolve in the interior of  $X_{\text{ad}}$  and only reach the boundary on convergence. We shall describe here one of these algorithms that we also call the **central trajectory algorithm**. There are two new ideas in this method: first, we penalize certain constraints with the help of potentials or ‘barrier’ functions; second, we use a Newton method to go from one iterate to the following.

Let us describe this method for the standard linear programming problem

$$\inf_{x \in \mathbb{R}^n \text{ such that } Ax=b, x \geq 0} c \cdot x. \quad (11.9)$$

We define a logarithmic potential for  $x > 0$

$$\pi(x) = - \sum_{i=1}^n \log x_i. \quad (11.10)$$

For a penalization parameter  $\mu > 0$ , we introduce the strictly convex problem

$$\min_{x \in \mathbb{R}^n \text{ such that } Ax=b, x>0} \mu\pi(x) + c \cdot x. \quad (11.11)$$

Let us remark that in practice the constraint  $x > 0$  is not important as it is never active: when we minimize (11.11) we cannot ‘approach’ the boundary of  $x > 0$  as we would ‘explode’ the potential  $\pi(x)$  to  $+\infty$ .

The principle of the central trajectory algorithm is to minimize (11.11) by a Newton method for smaller and smaller values of  $\mu$ . In effect, when  $\mu$  tends to zero, the penalized problem (11.11) tends to the linear programming problem (11.9).

**Exercise 11.2.6** Show that, if  $X_{\text{ad}}$  is bounded and nonempty, (11.11) has a unique optimal solution  $x^\mu$ . Write the optimality conditions and deduce that, if (11.9) has a unique optimal solution  $x^0$ , then  $x^\mu$  converges to  $x^0$  when  $\mu$  tends to zero.

### 11.2.4 Duality

The theory of duality (already stated in Section 10.3.3) is very useful in linear programming. Let us again consider the standard linear programming problem that we shall call primal (as opposed to dual)

$$\inf_{x \in \mathbb{R}^n \text{ such that } Ax=b, x \geq 0} c \cdot x, \quad (11.12)$$

where  $A$  is a matrix of size  $m \times n$ ,  $b \in \mathbb{R}^m$ , and  $c \in \mathbb{R}^n$ . For  $p \in \mathbb{R}^m$ , we introduce the Lagrangian of (11.12)

$$L(x, p) = c \cdot x + p \cdot (b - Ax), \quad (11.13)$$

where we have only ‘dualized’ the equality constraints. We introduce the associated dual function

$$G(p) = \min_{x \geq 0} L(x, p),$$

which, after calculation, becomes

$$G(p) = \begin{cases} p \cdot b & \text{if } A^*p - c \leq 0 \\ -\infty & \text{otherwise.} \end{cases} \quad (11.14)$$

The dual problem of (11.12) is therefore

$$\sup_{p \in \mathbb{R}^m \text{ such that } A^*p - c \leq 0} p \cdot b. \quad (11.15)$$

The space of admissible solutions of the dual problem (11.15) is denoted by

$$P_{\text{ad}} = \{p \in \mathbb{R}^m \text{ such that } A^*p - c \leq 0\}.$$

Let us recall that the space of admissible solutions of (11.12) is

$$X_{\text{ad}} = \{x \in \mathbb{R}^n \text{ such that } Ax = b, x \geq 0\}.$$

The linear programming problems (11.12) and (11.15) are called **duals**. The interest in this idea comes from the following result which is a particular case of the duality theorem 10.3.11.

**Theorem 11.2.11** *If (11.12) or (11.15) has an optimal finite value, then there exists  $\bar{x} \in X_{\text{ad}}$  an optimal solution of (11.12) and  $\bar{p} \in P_{\text{ad}}$  an optimal solution of (11.15) which satisfy*

$$\left( \min_{x \in \mathbb{R}^n \text{ such that } Ax=b, x \geq 0} c \cdot x \right) = c \cdot \bar{x} = \bar{p} \cdot b = \left( \max_{p \in \mathbb{R}^m \text{ such that } A^*p - c \leq 0} p \cdot b \right) \quad (11.16)$$

*Further,  $\bar{x}$  and  $\bar{p}$  are optimal solutions of (11.12) and (11.15) if and only if they satisfy the Kuhn–Tucker optimality conditions*

$$A\bar{x} = b, \quad \bar{x} \geq 0, \quad A^*\bar{p} - c \leq 0, \quad \bar{x} \cdot (c - A^*\bar{p}) = 0. \quad (11.17)$$

*If (11.12) or (11.15) has an optimal infinite value, then the set of admissible solutions of the other problem is empty.*

**Remark 11.2.12** An immediate consequence of the duality theorem 11.2.11 is that, if  $x \in X_{\text{ad}}$  and  $p \in P_{\text{ad}}$  are two admissible solutions of (11.12) and (11.15), respectively, they satisfy

$$c \cdot x \geq b \cdot p.$$

Likewise, if  $\bar{x} \in X_{\text{ad}}$  and  $\bar{p} \in P_{\text{ad}}$  satisfy

$$c \cdot \bar{x} = b \cdot \bar{p}$$

then  $\bar{x}$  is an optimal solution of (11.12) and  $\bar{p}$  of (11.15). These two properties allow us easily to find bounds for the optimal values of (11.12) and (11.15), and to test if a couple  $(\bar{x}, \bar{p})$  is optimal. •

**Proof.** Let us assume that  $X_{\text{ad}}$  and  $P_{\text{ad}}$  are nonempty. Take  $x \in X_{\text{ad}}$  and  $p \in P_{\text{ad}}$ . As  $x \geq 0$  and  $A^*p \leq c$ , we have

$$c \cdot x \geq A^*p \cdot x = p \cdot Ax = p \cdot b,$$

since  $Ax = b$ . In particular, this inequality implies that the optimal values of the two problems, primal and dual, are finite, therefore they have optimal solutions as a consequence of lemma 11.2.3. The equality (11.16) and the optimality condition (11.17) are then a consequence of the duality theorem 10.3.11.

Let us assume now that one of the two problems, primal or dual, has an optimal finite value. To fix ideas, let us assume that it is the dual problem (a symmetric



argument works for the primal problem). Then, lemma 11.2.3 tells us that there exists an optimal solution  $\bar{p}$  of (11.15). If  $X_{ad}$  is nonempty, we are reduced to the preceding situation which finishes the proof. Let us show therefore that  $X_{ad}$  is nonempty by again using the Farkas lemma 10.2.17. For  $p \in \mathbb{R}^m$ , we introduce the vectors of  $\mathbb{R}^{m+1}$

$$\tilde{b} = \begin{pmatrix} b \\ -b \cdot \bar{p} \end{pmatrix} \quad \text{and} \quad \tilde{p} = \begin{pmatrix} p \\ 1 \end{pmatrix}.$$

We verify that  $\tilde{b} \cdot \tilde{p} = b \cdot p - b \cdot \bar{p} \leq 0$ , for every  $p \in P_{ad}$ . On the other hand, the condition  $p \in P_{ad}$  can be rewritten

$$\tilde{p} \in C = \left\{ \tilde{p} \in \mathbb{R}^{m+1} \text{ such that } \tilde{p}_{m+1} = 1, \tilde{A}^* \tilde{p} \leq 0 \right\} \quad \text{with } \tilde{A} = \begin{pmatrix} A \\ -c^* \end{pmatrix}.$$

As  $\tilde{b} \cdot \tilde{p} \leq 0$  for every  $\tilde{p} \in C$ , the Farkas lemma 10.2.17 tells us that there exists  $\tilde{x} \in \mathbb{R}^n$  such that  $\tilde{x} \geq 0$  and  $\tilde{b} = \tilde{A} \tilde{x}$ , that is,  $\tilde{x} \in X_{ad}$  which is therefore not empty.

Finally, let us assume that the optimal value of the primal problem (11.12) is  $-\infty$ . If  $P_{ad}$  is nonempty, for every  $x \in X_{ad}$  and every  $p \in P_{ad}$ , we have  $c \cdot x \geq b \cdot p$ . By taking a minimizing sequence in  $X_{ad}$  we obtain  $b \cdot p = -\infty$ , which is absurd. Thus  $P_{ad}$  is empty. A similar argument shows that, if the optimal value of (11.12) is infinite, then  $X_{ad}$  is empty.  $\square$

The interest in duality to solve the linear programming problem (11.12) is multiple. On the one hand, depending on the algorithm chosen, it can be easier to solve the dual problem (11.15) (which has  $m$  variables and  $n$  inequality constraints) than the primal problem (11.12) (which has  $n$  variables,  $m$  equality constraints, and  $n$  inequality constraints). On the other hand, we can construct very efficient numerical algorithms for the solution of (11.12) which use the two forms, primal and dual, of the linear programming problem.

**Exercise 11.2.7** Use duality to solve 'by hand' (and without calculation!) the linear programming problem

$$\min_{x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0} 8x_1 + 9x_2 + 4x_3 + 6x_4$$

under the constraints

$$\begin{cases} 4x_1 + x_2 + x_3 + 2x_4 \geq 1 \\ x_1 + 3x_2 + 2x_3 + x_4 \geq 1 \end{cases}$$

**Exercise 11.2.8** Find the dual problem of (11.12) when we also dualize the constraint  $x \geq 0$ , that is, when we introduce the Lagrangian

$$L(x, p, q) = c \cdot x + p \cdot (b - Ax) - q \cdot x$$

with  $q \in \mathbb{R}^n$  such that  $q \geq 0$ . Compare with (11.15) and interpret the new dual variable  $q$ . Deduce from this that there is no interest in also 'dualizing' the constraint  $x \geq 0$ .

**Exercise 11.2.9** Verify that the dual problem of (11.15) is again (11.12).

**Exercise 11.2.10** Take  $v \in \mathbb{R}^n$ ,  $c \in \mathbb{R}^n$ ,  $A$  a matrix of size  $m \times n$  and  $b \in \mathbb{R}^m$ . We consider the linear programming problem

$$\inf_{\substack{v \geq 0 \\ Av \leq b}} c \cdot v. \quad (11.18)$$

Show that the dual problem can be put in the following form, with  $q \in \mathbb{R}^m$

$$\sup_{\substack{q \geq 0 \\ A^*q \leq c}} b \cdot q. \quad (11.19)$$

Let  $v$  and  $q$  be admissible solutions of (11.18) and (11.19), respectively. Show that  $v$  and  $q$  are optimal solutions if, and only if,

$$(c - A^*q) \cdot v = 0 \quad \text{and} \quad (b - Ac) \cdot q = 0. \quad (11.20)$$

The two equalities of (11.20) are called **complementary slack conditions** (primal and dual, respectively). Generalize to the case where the primal problem also includes equality constraints.

## 11.3 Integer polyhedra

We have until now treated problems of **continuous optimization**: the function to be minimized was differentiable, and the set of admissible solutions was defined by the intersection of a finite number of inequality constraints, likewise differentiable. **Combinatorial optimization**, conversely, treats problems for which the set of admissible solutions is **discrete**. Thus, in the case of the assignment problem which was the object of example 9.1.2, the set of admissible solutions was the set of permutations of  $n$  elements. The difficulty of combinatorial problems is on the one hand, that we cannot enumerate the set of admissible solutions, which is too large (with cardinality  $n!$  in the case of the assignment problem), and on the other hand, that the discrete nature of the space of solutions does not allow us to write optimality conditions directly with the help of the differential calculus.

We shall, however, see in the rest of this chapter that, despite appearances, the methods of continuous optimization are useful in combinatorial optimization. Let us consider the typical combinatorial problem

$$\sup_{x \in P \cap \mathbb{Z}^n} c \cdot x, \quad (11.21)$$

where  $P$  is a polyhedron of  $\mathbb{R}^n$ , that is, an intersection of a finite number of half-spaces, and  $c \in \mathbb{R}^n$ . The formulation (11.21) shows the difference between a combinatorial problem and a continuous problem: if we ignore the integer constraint in (11.21), we obtain

$$\sup_{x \in P} c \cdot x, \quad (11.22)$$

which is a linear programming problem, sometimes qualified as a **relaxed** continuous version of (11.21). (Generally, we talk of a relaxed problem, when we ignore certain constraints.) Relaxed problem (11.22) can be effectively treated by the methods of the preceding sections: all the difficulty in (11.21) comes from the fact that we restrict ourselves to the integer points of the polyhedron  $P$  (by **integer point**, we mean a point with integer coordinates). We shall now characterize the case where the solution of the discrete problem (11.21) is equivalent to that of its continuous relaxed version (11.22). These cases, which can seem exceptional, are in fact very important in practice, as they appear naturally in a certain number of concrete combinatorial problems: shortest path, assignment, and more generally flow problems with minimum cost.

### 11.3.1 Extreme points of compact convex sets

The idea which will allow us to link combinatorial problems and discrete problems is that of **extremal point**, an idea already seen in definition 11.2.1: an extremal point of a convex set  $K$  is a point  $x$  such that  $x = (y + z)/2$  and  $y, z \in K$  implies  $y = z = x$ . We denote by  $\text{extr } K$  the set of extremal points of  $K$ . Let us recall also that if  $X$  is a subset of  $\mathbb{R}^n$ , we say **convex envelope** of  $X$ , and we say  $\text{co } X$ , for the smallest convex set containing  $X$ , which we verify is equal to the set of barycentres of a finite number of elements of  $X$ . The **closed convex envelope** of  $X$ , denoted by  $\overline{\text{co}} X$ , is the smallest convex closed set containing  $X$ . It is equal to the closure of  $\text{co } X$ . The following result is fundamental.

**Theorem 11.3.1 (Minkowski)** *A compact convex set of  $\mathbb{R}^n$  is the convex envelope of the set of its extremal points.*

This theorem confirms therefore that  $K = \text{co } \text{extr } K$  when  $K$  is a convex compact set of  $\mathbb{R}^n$ . *A fortiori*,  $K = \overline{\text{co}} \text{extr } K$  since  $K$  is closed. The proof of the Minkowski theorem relies on the idea of a hyperplane of support, introduced in the appendix on Hilbert spaces: an affine hyperplane  $H = \{y \in \mathbb{R}^n \mid c \cdot y = \alpha\}$ , with  $c \in \mathbb{R}^n$ ,  $c \neq 0$ , and  $\alpha \in \mathbb{R}$  is a **hyperplane of support** of a convex set  $K$ , at the point  $x \in K$ , if  $\alpha = c \cdot x \leq c \cdot y$ , for every  $y \in K$ . We will use the following observation.

**Lemma 11.3.2** *If  $H$  is a hyperplane of support of a convex set  $K \subset \mathbb{R}^n$ , then every extremal point of  $H \cap K$  is an extremal point of  $K$ .*

**Proof.** Take  $H = \{y \in \mathbb{R}^n \mid c \cdot y = \alpha\}$  with  $c \in \mathbb{R}^n$ ,  $c \neq 0$ , and  $\alpha \in \mathbb{R}$ , a hyperplane of support of  $K$ . If  $x = (y + z)/2$  with  $y, z \in K$ , and if  $x \in K \cap H$ , we have  $\alpha = c \cdot x = (c \cdot y + c \cdot z)/2$ , and as  $\alpha \leq c \cdot y$  and  $\alpha \leq c \cdot z$ , we must have  $\alpha = c \cdot y = c \cdot z$ , therefore  $y, z \in K \cap H$ . If we assume that  $x$  is an extremal point of  $K \cap H$ , we therefore have  $x = y = z$ , which shows that  $x$  is an extremal point of  $K$ .  $\square$

**Proof of the Minkowski theorem 11.3.1.** We suppose obviously that  $K \neq \emptyset$  (otherwise, the result is trivial). Let us recall that the dimension of a nonempty

convex set is by definition the dimension of the affine space which it generates. We shall prove the theorem by induction on the dimension of  $K$ . Replacing  $\mathbb{R}^n$  by an affine subspace, we can assume that  $K$  has dimension  $n$ . If  $n = 0$ ,  $K$  is reduced to a point, and the theorem is satisfied. Let us assume therefore that the theorem is proved for compact convex sets of dimension at most  $n - 1$ , and let us show that every point  $x$  of  $K$  is a barycentre of a finite number of extremal points of  $K$ . If  $x$  is a boundary point of  $K$ , corollary 12.1.20 gives a hyperplane of support  $H$  of  $K$  at  $x$ . As  $K \cap H$  is a compact convex set of dimension at most  $n - 1$ , by the induction hypothesis,  $x$  is a barycentre of a finite number of extremal points of  $K \cap H$ , which are also extremal points of  $K$  according to lemma 11.3.2. Let us now take an arbitrary point  $x$  of  $K$ , and let  $D$  be an affine line passing through  $x$ . The set  $D \cap K$  is a segment of the form  $[y, z]$ , where the points  $y, z$  are boundary points of  $K$ . From before,  $y$  and  $z$  are barycentres of a finite number of extremal points of  $K$ . As  $x$  is itself a barycentre of  $y$  and  $z$ , the theorem is proved.  $\square$

**Remark 11.3.3** The Minkowski theorem is a particular case in finite dimensions of a result from functional analysis, the Krein–Milman theorem, which says that a compact convex set is a closed convex envelope of the set of extremal points (this result, which is a consequence of the Hahn–Banach theorem, holds in very general spaces and in particular in the Banach spaces). We note that in infinite dimensions, it is the closed convex envelope, and not the convex envelope, which occurs in the statement of the theorem.  $\bullet$

We now apply the Minkowski theorem to the problem of combinatorial optimization (11.21). In this case, the cost function  $J(x) = c \cdot x$  is linear, but it will be clearer to consider more generally the **maximization** of convex functions, which has very different properties from the **minimization** of convex functions treated in Chapters 9 and 10. We also consider an arbitrary set  $X$ , instead of  $P \cap \mathbb{Z}^n$ .

**Proposition 11.3.4 (maximization of a convex function)** *For every convex function  $J: \mathbb{R}^n \rightarrow \mathbb{R}$ , and for every subset  $X \subset \mathbb{R}^n$ ,*

$$\sup_{x \in X} J(x) = \sup_{x \in \text{co } X} J(x) = \sup_{x \in \overline{\text{co}} X} J(x) , \quad (11.23)$$

*and if  $X$  is bounded,*

$$\sup_{x \in X} J(x) = \sup_{x \in \text{extr } \overline{\text{co}} X} J(x) . \quad (11.24)$$

**Proof.** If  $y \in \text{co } X$ , we can write  $y = \sum_{1 \leq i \leq k} \alpha_i x_i$ , with  $x_i \in X$ ,  $\alpha_i \geq 0$ , and  $\sum_{1 \leq j \leq k} \alpha_j = 1$ . Since  $J$  is convex, we have  $J(y) \leq \sum_{1 \leq j \leq k} \alpha_j J(x_j) \leq \max_{1 \leq j \leq k} J(x_j) \leq \sup_{x \in X} J(x)$ , and since this is true for every  $y \in \text{co } X$ , we have  $\sup_{x \in \text{co } X} J(x) \leq \sup_{x \in X} J(x)$ . In addition, for every  $z \in \overline{\text{co}} X$ , we can write  $z = \lim_{k \rightarrow \infty} y_k$ , with  $y_k \in \text{co } X$ . As a convex function  $\mathbb{R}^n \rightarrow \mathbb{R}$  must be continuous (cf. exercise 9.2.7) we have  $J(z) = \lim_{k \rightarrow \infty} J(y_k) \leq \sup_{x \in \text{co } X} J(x)$ , and since this is true for every  $z \in \overline{\text{co}} X$ , we have  $\sup_{x \in \overline{\text{co}} X} J(x) \leq \sup_{x \in \text{co } X} J(x)$ . The other inequalities being trivial, we have shown (11.23). When  $X$  is bounded,  $\overline{\text{co}} X$  which is also bounded,

is compact. From the Minkowski theorem 11.3.1,  $\overline{\text{co}} X = \text{co extr } \overline{\text{co}} X$ , and by applying (11.23),  $\sup_{x \in \text{extr } \overline{\text{co}} X} J(x) = \sup_{x \in \text{co extr } \overline{\text{co}} X} J(x) = \sup_{x \in \overline{\text{co}} X} J(x) = \sup_{x \in X} J(x)$ , which proves (11.24).  $\square$

Proposition 11.3.4 suggests to us to consider the convex envelope of the admissible set  $X = P \cap \mathbb{Z}^n$  of our initial problem (11.21).

**Definition 11.3.5** We say **integer envelope** of a polyhedron  $P \subset \mathbb{R}^n$ , for the convex envelope of the set of integer points of  $P$ , which we denote by  $P_e = \text{co}(P \cap \mathbb{Z}^n)$ .

The term ‘integer envelope’ is traditional but misleading: usually, an envelope is a larger object, whereas here  $P_e \subset P$ .

**Corollary 11.3.6** If  $J: \mathbb{R}^n \rightarrow \mathbb{R}$  is convex, and if  $P \subset \mathbb{R}^n$  is a polyhedron, then

$$\sup_{x \in P \cap \mathbb{Z}^n} J(x) = \sup_{x \in P_e} J(x). \quad (11.25)$$

Thus, we can always replace the discrete problem (11.21) by a problem whose admissible set is convex. When  $J$  is linear, the problem on the right of (11.25) is a classical linear programming problem: we have thus concentrated the difficulty in the calculation, or the approximation, of the polyhedron  $P_e$ . There is a case where everything becomes easy.

**Definition 11.3.7** We say that a polyhedron  $P$  is an **integer polyhedron** if  $P = P_e$ .

We shall now give sufficient (precise) conditions so that a polyhedron is integral.

### 11.3.2 Totally unimodular matrices

An arbitrary polyhedron can be written

$$P = \{x \in \mathbb{R}^n \mid Ax \leq b\} \quad (11.26)$$

with  $A \in \mathbb{R}^{m \times n}$  and  $b \in \mathbb{R}^m$ . It should be noted that such a polyhedron is more general than the polyhedron  $X_{\text{ad}}$  of admissible solutions of the standard linear programming problem (cf. definition 11.2.1): indeed,  $X_{\text{ad}}$  is by definition included in the positive cone of  $\mathbb{R}^n$ , and we have besides already noted that  $X_{\text{ad}}$ , if it is nonempty, always has extremal points which is not the case for an arbitrary polyhedron (cf. remark 11.2.7). The characterization of extremal points of  $X_{\text{ad}}$  (lemma 11.2.5) extends, however, in the following way.

**Lemma 11.3.8** An extremal point of the polyhedron  $P$  defined by (11.26) must be a solution of a system  $A'x = b'$ , where  $A'$  is an invertible submatrix composed of  $n$  rows of  $A$ , and  $b'$  is the vector composed of the corresponding rows of  $b$ .

**Proof.** Let  $x$  be an extremal point of  $P$ , and take  $I(x) = \{1 \leq i \leq m \mid A_i \cdot x = b_i\}$  (the set of active constraints at  $x$ ), where  $A_i$  denotes the  $i$ th row of  $A$ . If the family  $\{A_i\}_{i \in I(x)}$ , does not have rank  $n$ , we can find a nonzero vector  $y$  such that  $A_i \cdot y = 0$  for every  $i \in I(x)$ . As  $x$  is the middle of the points  $x - \epsilon y$  and  $x + \epsilon y$ , which are elements of  $P$  if  $\epsilon$  is small enough, we contradict the extremality of  $x$ . Thus, we can find a subset  $I' \subset I(x)$  of cardinality  $n$  such that the  $n \times n$  matrix whose rows are  $A_i$ , with  $i \in I'$ , is invertible. The system  $A_i \cdot x = b_i, i \in I'$ , then characterizes  $x$ .  $\square$

Lemma 11.3.8 shows in particular that a polyhedron only has a finite number of extremal points.

**Exercise 11.3.1** Show conversely that if  $x$  is a point of  $P$  satisfying  $A'x = b'$ , with  $A'$  and  $b'$  as in lemma 11.3.8, then  $x$  is an extremal point.

Lemma 11.3.8 suggests studying the case where the solution of a linear system is an integer.

**Proposition 11.3.9** *Let  $A \in \mathbb{Z}^{n \times n}$  be an invertible matrix. The following assertions are equivalent:*

- (1)  $\det A = \pm 1$ ;
- (2) *for every  $b \in \mathbb{Z}^n$ , we have  $A^{-1}b \in \mathbb{Z}^n$ .*

**Proof.** The implication (1) $\Rightarrow$ (2) follows immediately from Cramer's formulas. Conversely, let us assume that  $A$  satisfies assertion 11.3.9. Let us show first that  $A^{-1}$  has integer coefficients. By taking  $b$  as the  $i$ th vector of the canonical basis of  $\mathbb{R}^n$ , we see that the  $i$ th column of  $A^{-1}$ , which coincides with  $A^{-1}b$ , has integer coefficients. As this is true for every  $1 \leq i \leq n$ , we have  $A^{-1} \in \mathbb{Z}^{n \times n}$ . Thus  $\det A^{-1} \in \mathbb{Z}$ , and  $1 = \det A \det A^{-1}$  shows that  $\det A$  divides 1, that is,  $\det A = \pm 1$ .  $\square$

**Definition 11.3.10** *We say that a matrix  $A \in \mathbb{Z}^{n \times n}$  is **unimodular** when  $\det A = \pm 1$ , and that a matrix  $B \in \mathbb{Z}^{m \times n}$  is **totally unimodular** when every square submatrix extracted from  $B$  has determinant  $\pm 1$  or 0.*

By taking  $1 \times 1$  submatrices, we see in particular that the coefficients of a totally unimodular matrix must be  $\pm 1$  or 0. The introduction of totally unimodular matrices is justified by the following result.

**Corollary 11.3.11 (optimality of integer solutions)** *Let  $D \in \mathbb{Z}^{m \times n}$  be a totally unimodular matrix,  $f \in (\mathbb{Z} \cup \{+\infty\})^m$ ,  $f' \in (\mathbb{Z} \cup \{-\infty\})^m$ ,  $g \in (\mathbb{Z} \cup \{+\infty\})^n$ , and  $g' \in (\mathbb{Z} \cup \{-\infty\})^n$ . Then, the extremal points of the polyhedron*

$$Q = \{x \in \mathbb{R}^n \mid f' \leq Dx \leq f, \quad g' \leq x \leq g\} \quad (11.27)$$

must be integer. In particular, if  $Q$  is bounded, we have  $Q = Q_e$ , and for every convex function  $J$  of  $\mathbb{R}^n$  in  $\mathbb{R}$ , we have

$$\sup_{x \in Q} J(x) = \sup_{x \in Q \cap \mathbb{Z}^n} J(x) . \quad (11.28)$$

**Proof.** We can write

$$Q = \{x \in \mathbb{R}^n \mid Ax \leq b\} , \quad (11.29)$$

where  $b$  is a finite integer vector and  $A$  is a matrix where each row is either of the form  $\pm D_i$ , with  $D_i$  an arbitrary row of  $D$ , or of the form  $\pm e_j$ , where  $e_j$  is the  $j$ th vector of the canonical basis of  $\mathbb{R}^n$ , for an arbitrary index  $1 \leq j \leq n$ .

We first show that  $A$  is totally unimodular. Let  $M$  be a  $k \times k$  submatrix extracted from  $A$ . Let us show by induction over  $k$  that  $\det M \in \{\pm 1, 0\}$ . If  $k = 1$ , this follows immediately from the total unimodularity of  $D$ . Let us now assume that the result is proved for all square submatrices of  $A$  with dimension at most  $k - 1$ , and let us prove it for  $M$ . If  $M$  contains a row equal to a vector  $\pm e_j$ , we expand  $\det M$  with respect to this row, and by induction, the result is proved. If  $M$  contains two rows equal up to a given sign,  $\det M = 0$ , and the result is again proved. Otherwise,  $M$  coincides, up to a change of sign of some row, with a submatrix of  $D$ , and as  $D$  is totally unimodular,  $\det M \in \{\pm 1, 0\}$ , which finishes the proof of the total unimodularity of  $A$ .

As  $Q$  is given by (11.29), with  $b$  an integer and  $A$  totally unimodular, it follows from lemma 11.3.8 and from proposition 11.3.9 that the extremal points of  $Q$ , if they exist, are integer.

If we suppose besides that  $Q$  is bounded,  $Q$  is compact, and from the Minkowski theorem 11.3.1,  $Q = \text{coextr } Q$ . As  $\text{extr } Q$  is composed of integer vectors,  $Q_e = \text{co}(Q \cap \mathbb{Z}^n) \supset \text{coextr } Q = Q$ , and in addition the inclusion  $Q_e \subset Q$  is trivial. The equality (11.28) is then obtained by applying corollary 11.3.6.  $\square$

**Exercise 11.3.2** We shall establish the reciprocal of corollary 11.3.11. Let us start by examining the special case of the polyhedron  $X_{\text{ad}} = \{x \in \mathbb{R}^n \mid Ax = b, x \geq 0\}$  of admissible solutions of the standard linear programming problem, with  $A \in \mathbb{Z}^{m \times n}$  with range  $m$ . Show that the two following properties are equivalent:

- (1) for every  $b \in \mathbb{Z}^m$ , the extremal points of  $X_{\text{ad}}$  are integer;
- (2) all the  $m \times m$  submatrices of  $A$  have determinant  $\pm 1$  or 0.

Let now  $D \in \mathbb{Z}^{m \times n}$ , and let us consider  $Q = \{x \in \mathbb{R}^n \mid Dx \leq b, x \geq 0\}$ . Deduce from the equivalence above the equivalence of the following two properties (Hoffman–Kruskal theorem):

- (3) for every  $b \in \mathbb{Z}^m$ , the extremal points of  $Q$  are integer;
- (4)  $D$  is totally unimodular.

**Remark 11.3.12** We will note that corollary 11.3.11 does not set any condition on  $J$ , except the convexity. In particular, if  $J(x) = c \cdot x$  is linear, the integer character of the optimal solutions is not directly linked to the integer character of the cost vector  $c$ . •

**Remark 11.3.13** The corollary 11.3.11 does not say that all the optimal solutions are integer. Moreover, when  $J$  is linear, it cannot be thus unless the optimal solution is unique, as every barycentre of optimal solutions of a linear programming problem is an optimal solution. •

**Remark 11.3.14** The condition that  $Q$  is bounded is not necessary to ensure that  $Q = Q_c$  in corollary 11.3.11: we limit ourselves to bounded polyhedra, which are sufficient in practice to model most of the combinatorial problems, to simplify the exposition. See [37] for more detail. •

There exist numerous results on totally unimodular matrices. We restrict ourselves here to giving a very useful sufficient condition.

**Proposition 11.3.15 (Poincaré)** *If  $A$  is a matrix with coefficients  $\pm 1$  or  $0$ , with at most one coefficient  $1$  per column, and at most one coefficient  $-1$  per column, then  $A$  is totally unimodular.*

**Proof.** As the property that  $A$  satisfies pass to the submatrices, it is enough to see that if  $A$  is square,  $\det A \in \{\pm 1, 0\}$ . If  $A$  has a zero column,  $\det A = 0$ . If  $A$  has a column with only one nonzero coefficient, we expand the determinant with respect to this column, and we conclude by induction that  $\det A \in \{\pm 1, 0\}$ . It only remains to consider the case where each column of  $A$  has exactly one coefficient  $1$  and exactly one coefficient  $-1$ : then, each column has zero sum, therefore  $\det A = 0$ . □

We apply proposition 11.3.15 to flow problems in the following section. Let us give for the moment as an exercise a case where we can calculate the total unimodularity by hand.

**Exercise 11.3.3 (covering problem)** A telephone call centre has a load curve:  $c_t$  is the number of clients being serviced at the discrete moment  $t \in \{1, \dots, T\}$ . A certain number of client advisers respond to the calls. We simplify the problem by assuming that all the calls are of the same type. We shall assume that there are  $k$  possible shifts, the shift  $i$  being characterized by an interval  $[\alpha_i, \beta_i]$ , with  $1 \leq \alpha_i \leq \beta_i \leq T$ , which amounts to ignoring breaks. We denote by  $S_i$  the salary of a client adviser working from the moment  $\alpha_i$  to the moment  $\beta_i$ . We set  $u_{it} = 1$  if  $\alpha_i \leq t \leq \beta_i$ , and  $u_{it} = 0$  otherwise. Justify the problem

$$\inf_{x \in \mathbb{N}^k} \sum_{1 \leq i \leq k} x_i S_i \quad . \quad (11.30)$$

$$\sum_{1 \leq i \leq k} x_i u_{it} \geq c_t, \quad \forall 1 \leq t \leq T$$



Show that the set of admissible solutions of this problem can be written as the set of integer points of a polyhedron of the form (11.27), where the matrix  $D$  is a matrix of intervals, that is, a matrix with coefficients 0, 1 such that the 1s appear consecutively in a column. Show that a matrix of intervals is totally unimodular. Conclude.

### 11.3.3 Flow problems

Before defining flow problems, let us consider an oriented graph  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ :  $\mathcal{N}$  is the set of **nodes**, and  $\mathcal{A} \subset \mathcal{N} \times \mathcal{N}$  is the set of **arcs**. An arc going from node  $i$  to node  $j$  is thus noted  $(i, j)$ . We equip each arc  $(i, j) \in \mathcal{A}$  with a **capacity**  $u_{ij} \in \mathbb{R}_+ \cup \{+\infty\}$  and a **cost**  $c_{ij} \in \mathbb{R}$ . (The ‘ $u$ ’ in  $u_{ij}$  is for ‘upper bound’.) We also take, at each node of the graph, an entrant flow  $b_i \in \mathbb{R}$  (if  $b_i < 0$ , it is an exiting flow, algebraically). We say **flow** for a function  $x \in \mathbb{R}^{\mathcal{A}}$ ,  $(i, j) \mapsto x_{ij}$ , satisfying the **Kirchoff law**

$$b_i + \sum_{j \in \mathcal{N}, (j, i) \in \mathcal{A}} x_{ji} = \sum_{j \in \mathcal{N}, (i, j) \in \mathcal{A}} x_{ij}, \quad \forall i \in \mathcal{N}, \quad (11.31)$$

and the positivity constraint

$$0 \leq x_{ij}, \quad \forall (i, j) \in \mathcal{A}. \quad (11.32)$$

By summing (11.31), we see that a necessary condition for the existence of a flow is that the sum of the entrant flows is zero

$$\sum_{i \in \mathcal{N}} b_i = 0. \quad (11.33)$$

We always assume that the condition (11.33) is satisfied. A flow is called **admissible** if it satisfies the capacity constraints

$$x_{ij} \leq u_{ij}, \quad \forall (i, j) \in \mathcal{A}. \quad (11.34)$$

**Definition 11.3.16** *We say flow problem with minimum cost for the linear programming problem*

$$\min \sum_{(i, j) \in \mathcal{A}} c_{ij} x_{ij} \quad \text{under the constraints } (11.31), (11.32), (11.34). \quad (11.35)$$

The flow problem with minimum cost has many important subproblems, such as the transport problem of example 9.1.1, or the assignment problem of example 9.1.2.

**Exercise 11.3.4** Make explicit the flow problem with minimum cost corresponding to the transport problem of example 9.1.1. (We will draw the graph.)

A particular fundamental case of the flow problem with minimum cost is the problem of maximal flow, or more properly the **flow problem**, which only concerns

the capacities (and not the costs). It will be convenient to assume that  $\mathcal{G}$  has two distinct nodes,  $s$  and  $p$ , called respectively **source** and **sink**, such that  $s$  has no predecessor ( $\{i \in \mathcal{N} \mid (i, s) \in \mathcal{A}\} = \emptyset$ ), and  $p$  has no successor ( $\{i \in \mathcal{N} \mid (p, i) \in \mathcal{A}\} = \emptyset$ ). Let  $v \in \mathbb{R}_+$ . We say **admissible flow from  $s$  to  $p$  with value  $v$**  for a solution  $x$  of (11.31), (11.32), (11.34), with

$$b_i = \begin{cases} v & \text{if } i = s, \\ -v & \text{if } i = p, \\ 0 & \text{otherwise.} \end{cases}$$

**Definition 11.3.17** *The maximal flow problem consists of finding an admissible flow from  $s$  to  $p$  of maximal value.*

**Exercise 11.3.5** Show that the maximal flow problem is effectively a particular case of the flow problem with minimal cost. (Hint: add an arc to the graph appearing in the definition of the maximal flow problem.)

In practice, we often look for integer solutions of a flow problem: for example, for the transport problem of example 9.1.1, the goods to be delivered could be parcels, and to deliver a half-parcel is meaningless. It is therefore natural to ask if a flow problem with minimum cost automatically has optimal integer solutions. Before applying corollary 11.3.11, let us note that the Kirchoff law (11.31) can be written  $Ax = b$ , where the matrix  $A \in \mathbb{R}^{\mathcal{N} \times \mathcal{A}}$ , called the **incidence matrix** of  $\mathcal{G}$ , is defined by

$$A_{i,(j,k)} = \begin{cases} -1 & \text{if } i = k, \\ 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

The matrix  $A$  is well defined, except in the degenerate case where the graph has a loop, that is, an arc  $(j, k)$  such that  $j = k$ . A flow circulating on a loop has a contribution which is simple in the Kirchoff law (11.31), and there is no loss of generality to assume that the graph does not have a loop, which we shall do in the rest of the section.

We can now write the set of admissible flows

$$\{x \in \mathbb{R}^{\mathcal{A}} \mid Ax = b, 0 \leq x \leq u\} . \quad (11.36)$$

**Proposition 11.3.18** *The incidence matrix of a graph is totally unimodular.*

**Proof.** This is an immediate consequence of proposition 11.3.15. □

**Corollary 11.3.19 (optimality of integer flows)** *If the entrant flows  $b_i$  are integer, and if the capacities  $u_{ij}$  are integer or infinite, then, the extremal points of the set (11.36) of admissible solutions of a flow problem with minimal cost are integer. In particular, if the set of admissible flows is bounded and nonempty, there exists an optimal integer flow.*

**Proof.** This is an immediate consequence of corollary 11.3.11 and of proposition 11.3.18.  $\square$

**Exercise 11.3.6** Show that the covering problem (exercise 11.3.3) can be modelled by a flow problem with minimum cost, and thus recover the conclusion of exercise 11.3.3.

**Exercise 11.3.7** Take again the assignment problem, introduced in example 9.1.2. We always consider  $n$  boys and  $n$  girls, but here,  $a_{ij}$  is a real number which represents the happiness of the couple  $(i, j)$ , and we look for a permutation  $\sigma \in \mathcal{S}_n$ , the optimal solution of

$$\max_{\sigma \in \mathcal{S}_n} \sum_{1 \leq i \leq n} a_{i\sigma(i)} . \quad (11.37)$$

1. Show that this problem is equivalent to the integer linear problem

$$\max_{x \in \mathcal{B}_n \cap \mathbb{Z}^{n \times n}} \sum_{1 \leq i, j \leq n} a_{ij} x_{ij}, \quad (11.38)$$

where  $\mathcal{B}_n$  denotes the set of bistochastic matrices, that is, the set of real matrices  $x = (x_{ij})$  of size  $n \times n$  such that

$$\begin{aligned} \forall i \in \{1, \dots, n\}, \quad 1 &= \sum_{1 \leq k \leq n} x_{ik}, \\ \forall j \in \{1, \dots, n\}, \quad 1 &= \sum_{1 \leq k \leq n} x_{kj}, \\ \forall i, j \in \{1, \dots, n\}, \quad 0 &\leq x_{ij} \end{aligned}$$

2. Show that the problem (11.38) is an integer flow problem with minimum cost (we shall draw the graph). Deduce that the polyhedron  $\mathcal{B}_n$  is an integer. What can we conclude as to the difficulty of the assignment problem?

3. Deduce from above that every bistochastic matrix is the barycentre of a finite number of permutation matrices (this theorem is due to Birkhoff).

4. Deduce from the above that if at a ball, there are  $n$  boys and  $n$  girls, each boy having been presented to  $r$  girls, and each girl having been presented to  $r$  boys, with  $r \geq 1$ , it is possible to form  $n$  dance couples so that the dancers of each couple have already been presented to each other (this theorem is due to König).

**Exercise 11.3.8** This exercise presents an algorithm which is fundamental in the theory of the flows, due to Ford and Fulkerson. We consider the maximal flow problem from a source  $s$  to a sink  $p$  in a graph  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  (see definition 11.3.17). For simplicity, we shall assume that each of the capacities  $u_{ij}$  do not take the value  $+\infty$ . For every  $I, J \subset \mathcal{N}$ , and for every  $x = (x_{ij}) \in \mathbb{R}^{\mathcal{A}}$ , we set  $x(I, J) = \sum_{i \in I, j \in J, (i, j) \in \mathcal{A}} x_{ij}$ . We say that a partition of  $\mathcal{N}$  into two subsets  $S$  and  $\bar{S}$  is a *cut separating  $s$  from  $p$*  if  $s \in S$  and  $p \in \bar{S}$ , and we say that  $u(S, \bar{S})$  is the *capacity* of this cut.

1. Show that for every cut  $(S, \bar{S})$  separating  $s$  from  $p$ , and for every admissible flow  $x$  from  $s$  to  $p$  with value  $v$ ,

$$x(S, \bar{S}) - x(\bar{S}, S) = v .$$

Deduce that the value of every admissible flow from  $s$  to  $p$  is bounded above by the capacity of each cut separating  $s$  from  $p$ .

2. Given an admissible flow  $x$  from  $s$  to  $p$ , we define the *residual graph*  $\mathcal{G}_r(x) = (\mathcal{N}, \mathcal{A}_r(x))$  where  $\mathcal{A}_r(x)$  denotes the set of couples  $(i, j)$  such that either  $(i, j) \in \mathcal{A}$  and  $x_{ij} < u_{ij}$ , or  $(j, i) \in \mathcal{A}$  and  $x_{ji} > 0$ . Show that if there exists a path  $\gamma$  from  $s$  to  $p$  in the residual graph  $\mathcal{G}_r(x)$ , it is possible to construct a new admissible flow  $x'$  from  $s$  to  $p$  with value strictly greater than  $x$ , by modifying only the values  $x_{ij}$  when  $(i, j)$  or  $(j, i)$  is an arc of the path  $\gamma$ . Notice besides that if  $x$  is integer and if the capacities are integer, we can choose  $x'$  integer.

3. We suppose now that there is no path from  $s$  to  $p$  in  $\mathcal{G}_r(x)$ . Let  $S$  be the set of accessible nodes since  $s$  is in  $\mathcal{G}_r(x)$ , and  $\bar{S}$  the complement of  $S$  is in  $\mathcal{N}$ . Show that  $(S, \bar{S})$  is a cut separating  $s$  from  $p$  whose capacity is equal to the value of the flow  $x$ .

4. Conclude that the maximal value of the flow from  $s$  to  $p$  is equal to the minimal capacity of a cut separating  $s$  from  $p$ . (This is the ‘max-flow min-cut’ theorem of Ford and Fulkerson.)

5. Deduce an algorithm allowing us, when the capacities are integers, to calculate a maximal flow from  $s$  to  $p$  in a time  $O(v^*|\mathcal{A}|)$ , where  $v^*$  is the maximal value of a flow from  $s$  to  $p$ , and  $|\mathcal{A}|$  is the number of arcs.

**Remark 11.3.20** We recognize in the ‘max-flow min-cut’ theorem of Ford and Fulkerson (Question 4 of exercise 11.3.8) a particular case of the duality theorem in linear programming. The reader will be able to write the dual of the problem of maximal value flow, and thus recover the Ford–Fulkerson theorem. •

**Remark 11.3.21** The Ford–Fulkerson algorithm, presented in exercise 11.3.8, is only the simplest of the flow algorithms. A variant of this algorithm, the Edmonds–Karp algorithm, (also due to Dinits), can be implemented in a time  $O(nm^2)$  independent of the integer character of the capacities, where  $n = |\mathcal{N}|$  denotes the number of nodes, and  $m = |\mathcal{A}|$  denotes the number of arcs. This refinement consists very simply of selecting, at each step of the Ford–Fulkerson algorithm, the shortest path from  $s$  to  $p$ , that is, the path from  $s$  to  $p$  which has the smallest number of arcs. There also exists a very different flow algorithm, the ‘preflow-push’ algorithm of Goldberg and Tarjan (1986), which has a time of execution of the order of  $O(n^2m)$ . All these ideas generalize to the case of the flow problem with minimum cost. See [1] for the state of the art. •

## 11.4 Dynamic programming and shortest path problems

Dynamic programming, developed by R. Bellman in the 1950s, is a very general method which is applied to decision problems in time (such as optimal control,

except that the time variable is sometimes disguised). For each problem the question is that of identifying a good idea of a **state**. With each state we associate an optimal value starting from this state, and the dynamic programming equation links the value of a state at a given moment to those states which are available at the following moment. (Dynamic programming is exactly to optimization what the Markovian point of view is to probability theory.)

### 11.4.1 Bellman's optimality principle

**Dynamic programming** is based on the Bellman optimality principle, which can be stated very simply in the particular case of the shortest path problem: if the shortest path from a town  $A$  to a town  $B$  passes through a town  $C$ , then the subpath going from  $A$  to  $C$  is again the shortest path from  $A$  to  $C$ . Denoting by  $d_{XY}$  the distance from  $X$  to  $Y$ , we obtain

$$d_{AB} = \min_C (d_{AC} + d_{CB}), \quad (11.39)$$

where the min is taken on the set of the towns  $C$  through which we can pass going from  $A$  to  $B$ . The equation (11.39) is a particular case of a dynamic programming equation, or the **Bellman equation**, it will allow us to calculate  $d$  recursively knowing that  $d_{XY}$  is known when  $X$  and  $Y$  are neighbouring towns. The idea of a state appears here naturally: we set ourselves the problem of calculating  $d_{AB}$ , where  $A$  and  $B$  are two fixed towns, and we see that it is useful to tabulate the distances  $d_{CB}$  (or  $d_{AC}$ ) for all the intermediate towns  $C$ . The details of the calculation must of course be fixed to give a true algorithm: it is this that we shall do in the two subsections which follow. First of all we treat a simpler version of the problem, where the time appears explicitly, in Section 11.4.2, then return to the shortest path problem, under a more general form, in Section 11.4.3.

### 11.4.2 Finite horizon problem

Let us consider a small problem of an economic nature, which merits the name of a problem of discrete optimal control at finite horizon. Let  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  be an oriented graph, equipped with weights  $c : \mathcal{A} \rightarrow \mathbb{R}$ , which we interpret as a **cost**. Let us recall that a **path** is a sequence of nodes linked consecutively by arcs, and that a path whose first and last node coincide is called a **circuit**. The cost of a path is by definition the sum of the costs of its arcs. Let us fix an integer  $T$  (the horizon), and an initial node  $i \in \mathcal{N}$ . We want to calculate the total cost starting from  $i$  to the horizon  $T$  that we denote

$$v_i^T = \min_{\substack{(\ell_0, \dots, \ell_T) \\ \ell_0 = i}} \text{path} \quad c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T}. \quad (11.40)$$

The function  $i \mapsto v_i^T$ , which we can represent by a vector  $v^T \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$ , is traditionally called the **function value** (at horizon  $T$ ). We agree by writing (11.40) that  $\min \emptyset = +\infty$ , which reduces to writing  $v_i^T = +\infty$  when there are no paths of length  $T$  starting from  $i$  in the graph. Another special case is  $T = 0$ : the paths which

appear in (11.40) are then of length zero. We agree that for every vertex of the graph, there is a circuit of length zero passing through this vertex, and their cost is zero. In particular,  $v_i^0 = 0$ .

A naive evaluation of (11.40) consists of enumerating all the paths of length  $T$ , whose number increases exponentially with  $T$ . Dynamic programming will allow us to factorize this calculation, and thus carry it out in polynomial time (the notion of polynomial time is defined in remark 11.6.6).

The idea for calculating  $v_i^T$  is to vary the horizon and the initial state, by calculating  $v_k^t$  for every  $0 \leq t \leq T$  and  $k \in \mathcal{N}$ . In effect, we can write for every  $i \in \mathcal{N}$ , and  $t \geq 1$ ,

$$v_i^t = \min_{k \in \mathcal{N}, (i,k) \in \mathcal{A}} (c_{i,k} + v_k^{t-1}) , \quad (11.41)$$

$$v_i^0 = 0. \quad (11.42)$$

The equation (11.41) follows from the Bellman optimality principle: the optimal cost starting from  $i$ , in  $t$  steps, is obtained by choosing the first movement  $i \rightarrow k$ , in such a way as to minimize the cost of this first movement, that is,  $c_{i,k}$  plus the optimal cost starting from  $k$  with horizon  $t - 1$ . The initial condition (11.42) is trivial: if nothing remains to be done we pay nothing. We shall write the dynamic programming equation, or **Bellman equation**, adapted to our problem: it allows us to calculate by induction the sequence of vectors  $v^0, v^1, \dots \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$ .

It is easy to solve more general problems by modifying (11.42). Let us consider, for example, the new value function

$$v_i^T = \min_{\substack{(\ell_0, \dots, \ell_T) \text{ path} \\ \ell_0 = i}} c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T} + \phi_{\ell_T}, \quad (11.43)$$

where  $\phi \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$  is a vector representing a penalty associated with the final state. When  $\phi = 0$ , we recover (11.40). When  $\phi$  is the indicator function of a vertex  $j \in \mathcal{N}$ , that is,

$$\phi_k = \begin{cases} 0 & \text{if } k = j \\ +\infty & \text{otherwise,} \end{cases} \quad (11.44)$$

(11.43) forces us to finish in the state  $j$ , and  $v_i^T$  then gives the minimal cost in  $T$  steps to go from  $i$  to  $j$ . The function value (11.43) always satisfies (11.41), with the new initial condition for the Bellman equation

$$v^0 = \phi . \quad (11.45)$$

This is the moment to remark that the time  $T$  which occurs in the Bellman equation (11.41), (11.45) flows **in an inverse sense** from the physical time which occurs in the trajectories (11.43): in our modelling,  $T$  is the time which remains, also a penalty on the **terminal** state of the trajectory  $(\ell_0, \dots, \ell_T)$  leads to changing the **initial** condition of the Bellman equation. We say that the Bellman equation is a **retrograde** equation. This inversion of time is inherent in decision problems: it is

often only at a bad end that we understand that we did not need to play at the beginning.

The interest in the Bellman equation (11.41) is that  $v^t$  can be calculated starting from  $v^{t-1}$  in time  $O(|\mathcal{A}| + |\mathcal{N}|)$ , where  $|\mathcal{A}|$  and  $|\mathcal{N}|$  denote respectively the number of arcs and the number of nodes, see remark 11.4.1 for more detail. Thus,  $v^T$  can be calculated in time  $O(T(|\mathcal{A}| + |\mathcal{N}|))$ , compared with, for example, the time  $O(|\mathcal{N}|p^T)$  of a naive algorithm enumerating the paths of length  $T$  in a graph where each vertex has exactly  $p$  successors. We very simply obtain, on the other hand, an optimal path from the Bellman equation (11.41), (11.45): we set  $\ell_0 = i$ , then we choose  $\ell_1$  realizing the minimum in (11.41), that is,  $v_{\ell_0}^T = c_{\ell_0, \ell_1} + v_{\ell_1}^{T-1}$ , and more generally  $\ell_{r+1}$  such that  $v_{\ell_r}^{T-r} = c_{\ell_r, \ell_{r+1}} + v_{\ell_{r+1}}^{T-r-1}$ . By construction,  $v_i^T = c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T} + \phi_{\ell_T}$ , which shows that  $(\ell_0, \dots, \ell_T)$  is an optimal path for the problem (11.43).

**Example 11.4.1** To illustrate the above, let us consider the case of a taxi driver cruising in the imaginary town represented in Figure 11.2. The town is formed of three zones,  $H$ , a posh district,  $A$  an airport, and  $B$  a suburb (from where the driver usually returns empty). The taxi driver wants to make  $T$  runs, starting from  $H$ , and maximize the total gain, which is the sum of the gains of the runs  $c$ , and of a bonus,  $\phi$ , expressing a preference for the final state of the last run. We have represented the gains of the runs on the arcs, and the bonuses by the exiting arcs. The bonus  $-\infty$  at  $B$  means that the taxi does not want to finish its journey at  $B$ . We will assume in addition that the taxi has the ability to choose its runs. As this is a maximization problem, the dynamic programming equation is written with max instead of min, and  $-\infty$  instead of  $+\infty$ :

$$\begin{aligned} v_H^t &= \max(3 + v_H^{t-1}, 10 + v_A^{t-1}) \\ v_A^t &= \max(12 + v_H^{t-1}, 7 + v_B^{t-1}) \\ v_B^t &= \max(-5 + v_H^{t-1}, -3 + v_A^{t-1}) \end{aligned}$$

$$v_H^0 = 0, \quad v_A^0 = 2, \quad v_B^0 = -\infty .$$

Let us now calculate the optimal strategy of the taxi at horizon 2 starting from  $H$ . It is enough to evaluate

$$v_H^1 = \max(3 + 0, \underline{10 + 2}) = 12 \quad (11.46)$$

$$v_A^1 = \max(\underline{12 + 0}, 7 + -\infty) = 12 \quad (11.47)$$

$$v_B^1 = \max(-5 + 0, \underline{-3 + 2}) = -1 \quad (11.48)$$

$$v_H^2 = \max(3 + 12, \underline{10 + 12}) = 22, \quad (11.49)$$

where we have underlined the terms which realize the maximum. We deduce that the optimal gain, 22, is obtained by first realizing the max in (11.49) (we go first from  $H$  to  $A$ ), then realizing the max in (11.47) (we return from  $A$  to  $H$ ). •

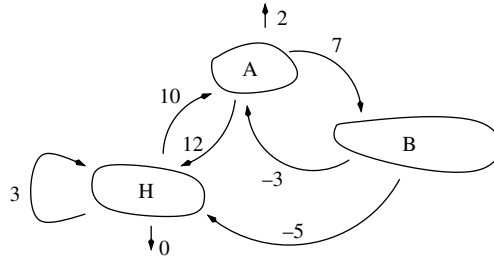


Figure 11.2. A very deterministic taxi driver.

**Remark 11.4.1** A classical way to code a graph  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  with  $n = |\mathcal{N}|$  nodes and  $m = |\mathcal{A}|$  arcs equipped with cost  $c$ , consists of defining three vectors: a vector  $\mathbf{h} \in \{1, \dots, n\}^m$  (for ‘head’), a vector  $\mathbf{t} \in \{1, \dots, n\}^m$  (for ‘tail’), as well as a vector of real numbers,  $\mathbf{c} \in \mathbb{R}^m$ . This representation reduces to numbering the arcs of  $\mathcal{A}$  from 1 to  $m$ , by saying that the arc  $(i, j) \in \mathcal{A}$  which has the number  $k$  going from  $\mathbf{h}_k = i$  to  $\mathbf{t}_k = j$ , and has a cost  $\mathbf{c}_k = c_{i,j}$ . The graph therefore occupies a space in memory of  $O(n + m)$ . We easily see that with the graph coded like this, it is possible to calculate  $v^t$  starting from  $v^{t-1}$  by using (11.41), in time  $O(n + m)$ . •

### 11.4.3 Minimum cost path, or optimal stopping, problem

Let us now consider the problem of the **minimum cost path** which, given two vertices  $i$  and  $j$ , consists of finding a path of arbitrary length going from  $i$  to  $j$  and with minimum cost. This is a generalization of the shortest path problem already stated in Section 11.4.1: as opposed to the case of distances, we do not assume here that the costs are positive. By exploiting the preceding notation, we must now calculate

$$v_i = \inf_{T \in \mathbb{N}} v_i^T = \inf_{\substack{(\ell_0, \dots, \ell_T) \text{ path} \\ T \in \mathbb{N}, \ell_0 = i}} c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T} + \phi_{\ell_T}, \quad (11.50)$$

where the vector  $\phi$  penalizing a final state other than  $j$  is given by (11.44). We can also consider an arbitrary penalty  $\phi \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$ , which does not change the results. We will note that  $v_i$  now has values in  $\mathbb{R} = \mathbb{R} \cup \{\pm\infty\}$  (the sums on the right of (11.50) can be  $+\infty$ , in addition, it is possible that these sums are not bounded below, as we consider for arbitrarily long paths). The function value  $v$  satisfies the new Bellman equation

$$v_i = \min(\phi_i, \min_{k \in \mathcal{N}, (i,k) \in \mathcal{A}} (c_{i,k} + v_k)), \quad \forall i \in \mathcal{N}. \quad (11.51)$$

By comparison with (11.41), the presence of a supplementary term in the min gives the possibility of stopping in any state  $k$  with the final penalty  $\phi_k$ , which in the special case (11.44), prevents us from stopping elsewhere than  $j$ .



We must now show that the system (11.51) allows us to calculate the function value. Let us recall that we call a **maximal solution** of an equation a solution which bounds all the others above.

**Theorem 11.4.2** *The function value  $v$  defined by (11.50) is the maximal solution of the Bellman equation (11.51).*

**Proof.** We have already shown that  $v$  satisfies (11.51). Let  $v' \in \overline{\mathbb{R}}^{\mathcal{N}}$  be an arbitrary solution of (11.51). Let us show that  $v' \leq v$ . Let  $(\ell_0, \dots, \ell_T)$  be an arbitrary path starting from  $i$ . As  $v'$  satisfies (11.51), we can write  $v'_{\ell_r} \leq c_{\ell_r, \ell_{r+1}} + v'_{\ell_{r+1}}$ , for every  $0 \leq r \leq T-1$ , and  $v'_{\ell_T} \leq \phi_{\ell_T}$ . By combining these inequalities,

$$v'_i \leq c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T} + \phi_{\ell_T}$$

and by taking the infimum over all the paths  $(\ell_0, \dots, \ell_T)$  starting from  $i$ ,  $v'_i \leq v_i$ . Thus  $v' \leq v$ , which shows that  $v$  is a maximal solution of (11.51).  $\square$

**Exercise 11.4.1** Show that the Bellman equation (11.51) can have many finite solutions. (Hint: consider a graph with a single vertex.)

**Exercise 11.4.2** We say that the graph is coaccessible for  $\phi$  if for every node  $i$  of the graph, there exists a path from  $i$  to a node  $j$  such that  $\phi_j \neq +\infty$ . We shall show that if the graph does not have a circuit of negative cost and is coaccessible for  $\phi$ , then the Bellman equation (11.51) has a unique finite solution, equal to  $v$ . For this, we consider  $v' \in \mathbb{R}^{\mathcal{N}}$  a solution of the Bellman equation (11.51). We introduce the continuation set

$$C = \{i \in \mathcal{N} \mid \phi_i \geq \min_{k \in \mathcal{N}, (i,k) \in \mathcal{A}} (c_{i,k} + v'_k)\},$$

and we choose for each  $i \in C$  a node  $\pi(i)$  such that

$$v'_i = c_{i, \pi(i)} + v'_{\pi(i)}.$$

We thus define a mapping  $\pi: C \rightarrow \mathcal{N}$ . Show that for arbitrary  $i \in C$ , there exists an integer  $k$  such that the  $k$ th iterate  $\pi^k(i)$ , does not belong to  $C$ . Conclude that  $v' \geq v$ .

We can rewrite (11.51) as a fixed point equation  $v = f(v)$ , with an obvious definition of  $f: \overline{\mathbb{R}}^{\mathcal{N}} \rightarrow \overline{\mathbb{R}}^{\mathcal{N}}$ . Also theorem 11.4.2 suggests calculating  $v$  with the help of a fixed point algorithm. To decide about the convergence of fixed point methods, we often refer to contraction arguments. Here, the convergence analysis will rather use the ordering. We equip  $\overline{\mathbb{R}}^{\mathcal{N}}$  with the usual partial order, defined component by component, and for this order, the mapping  $f: x \mapsto f(x)$  is increasing. Since  $f(\phi) \leq \phi$ , an immediate recurrence shows that  $f^{r+1}(\phi) \leq f^r(\phi)$ , for every  $r \geq 0$ . The sequence  $\{f^r(\phi)\}_{r \geq 0}$  which decreases must converge to a vector  $v' \in \overline{\mathbb{R}}^{\mathcal{N}}$ . (The reader will note here that we allow the value  $-\infty$  for the coefficients of  $v'$ .) We have  $f(v') = v'$  by continuity of  $f$ . On the other hand, if  $v''$  is an arbitrary fixed point of  $f$ ,

we have trivially  $v'' \leq \phi$ , therefore since  $f$  is increasing,  $v'' = f^r(v'') \leq f^r(\phi)$ , for every  $r \geq 0$ , therefore by taking the infimum on the  $r \geq 0$ ,  $v'' \leq v'$ , therefore  $v'$  is the largest fixed point of  $f$ . Since from theorem 11.4.2, the function value is also the largest fixed point of  $f$ , we shall show the following result.

**Theorem 11.4.3 (value iteration)** *The sequence  $\{f^r(\phi)\}_{r \geq 0}$  decreases to the function value  $v$  defined by (11.50).*

More generally, given a dynamic programming operator  $f$  of which we want to calculate a fixed point, the algorithm consists of constructing a sequence  $f^r(\phi)$ , either starting from an arbitrary supersolution  $\phi$  (we say that  $\phi$  is a supersolution of the fixed point equation  $x = f(x)$  if  $f(\phi) \leq \phi$ ), or starting from an arbitrary subsolution  $\phi$  (defined by reversing the inequality) is called **value iteration**. For general dynamic programming problems, particularly in stochastic control, a Newton type method, called policy iteration (see, for example, D. Bertsekas, *Dynamic Programming and Optimal Control*, Vol. I and II, Athena Sci., Belmont, MA, 1995), is often more rapid. The interest in value iteration is its simplicity. In the deterministic case, we have convergence in finite time for value iteration.

**Exercise 11.4.3 (convergence in finite time)** Show that if  $\mathcal{G}$  does not have circuits of strictly negative cost, then  $f^{|\mathcal{N}|-1}(\phi) = v$ . If conversely  $\mathcal{G}$  has a circuit of strictly negative cost, and if  $\mathcal{G}$  is coaccessible for  $\phi$  (see exercise 11.4.2 for this idea) then  $f^{|\mathcal{N}|}(\phi) < f^{|\mathcal{N}|-1}(\phi)$ .

Exercise 11.4.3 suggests implementing the value iteration algorithm as follows. We calculate by induction the sequence by satisfying at each step if  $x_r = x_{r-1}$ , in which case  $x_r = v$  and we stop (the interest in this test is that the time of convergence is often much smaller than  $|\mathcal{N}| - 1$ ). In the worst case, we arrive at  $r = |\mathcal{N}|$  and we stop then: we know that there exists a circuit of strictly negative cost. In practice, we rarely program this value iteration algorithm, but rather the following variant of Gauss–Seidel, called the Ford–Bellman algorithm, which updates as soon as possible all the coordinates in the value iteration. It is faster to program this variant than to describe it.

**Algorithm 11.4.4 (Ford–Bellman)** *Take  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  and  $c : \mathcal{A} \rightarrow \mathbb{R}$ , with variables:  $v \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$ ,  $b$  Boolean,  $r$  integer,  $i, k \in \mathcal{N}$ .*

*Initialization:  $r \leftarrow 0$ ,  $b \leftarrow \text{true}$ ; for every  $i \in \mathcal{N}$ ,  $v_i \leftarrow \phi_i$ .*

*As long as  $r < |\mathcal{N}|$  and  $b$  make:*

*$b \leftarrow \text{false}$ ;*

*$r \leftarrow r + 1$*

*for every  $i \in \mathcal{N}$  and for every  $(i, k) \in \mathcal{A}$ , if  $c_{i,k} + v_k < v_i$ , let  $v_i \leftarrow c_{i,k} + v_k$  and  $b \leftarrow \text{true}$ .*

*If  $r < |\mathcal{N}|$ ,  $v$  is the solution, if  $r = |\mathcal{N}|$ , there exists a circuit of strictly negative cost.*

In order to illustrate the value iteration, let us revisit the problem of the taxi driver represented in Figure 11.2, and we will be interested in the value of the paths of **maximum** gain and arbitrary length starting from  $H$ . As there is a circuit of gain which is strictly positive (going from  $H$  to  $A$ , and returning, which pays 22), our driver may find it beneficial to make an infinite number of runs, and we have in particular  $v_H = +\infty$ . In order to tackle

a problem which is less degenerate, let us, for example, take a tax of 11 from the price of each run. The operator  $f$  defined starting from (11.51) becomes

$$\begin{aligned} [f(x)]_H &= \max(0, -8 + x_H, -1 + x_A) \\ [f(x)]_A &= \max(2, 1 + x_H, -4 + x_B) \\ [f(x)]_B &= \max(-16 + x_H, -14 + x_A) . \end{aligned}$$

The value iteration consists of calculating  $x^0 = \phi$ ,  $x^1 = f(x^0)$ ,  $\dots$ , let

$$\begin{aligned} x^0 &= \phi = (0, 2, +\infty) \\ x_H^1 &= \max(0, -8 + 0, \underline{-1 + 2}) = 1 \\ x_A^1 &= \max(2, 1 + 0, -4 + -\infty) = 2 \\ x_B^1 &= \max(-16 + 0, \underline{-14 + 2}) = -12 \\ x^2 &= x^1 = v, \text{ stop,} \end{aligned}$$

and by considering the  $\arg \max$ , we see that it is optimal to stop when we are at  $A$ , and if we are at  $B$  or  $H$ , we must go to  $A$ .

**Exercise 11.4.4** How can we complete the Ford–Bellman algorithm to construct a circuit of negative cost?

**Remark 11.4.5** There exist many variants of the Ford–Bellman algorithm, which differ in the order in which we traverse the nodes  $i$  and the arcs  $(i, k)$  in the loop of the algorithm 11.4.4: we will find the generic term ‘label correcting algorithms’ in the literature. The algorithms in this family, and in particular the Ford–Bellman algorithm, are among the quickest for calculating paths of minimum cost, in the case of an oriented graph which has circuits (of positive or zero cost) and whose costs can be negative. In two special cases, we can do much better than Ford–Bellman. When the graph  $\mathcal{G}$  does not have a circuit, we make a **topological order**, that is, we equip the vertices with a total order  $\leq$ , such that if there is a path from  $i$  to  $j$ , then  $i \leq j$ . Often, the nodes are already naturally ordered (for example, by increasing time) and if they are not, we can find such an order in linear time; see, for example, [1] for more detail. (We say that an algorithm is in **linear time** if the time of execution is bounded by a constant times the size of the data, which is the best we can hope, see remark 11.6.6 for more detail on the idea of calculation time.) Once the nodes are ordered, it is enough to start  $v$  at  $+\infty$ , and to make once and only once the substitution  $v_i \leftarrow \min(\phi_i, \min_{k \in \mathcal{N}, (i,k) \in \mathcal{A}} c_{i,k} + v_k)$  for each  $i$ , by traversing the  $i$  in decreasing order, to obtain the function value, which takes a linear time. Another particular remarkable case is that where the costs are positive or zero: in this case, we can employ a (greedy) algorithm (cf. section 11.5), called the Moore–Dijkstra algorithm, see, for example, [1]. •

**Exercise 11.4.5** We return to example 9.1.3, that is the classical knapsack problem. We assume here that the weights are integer. We therefore take  $n$  objects with respective weights  $p_1, \dots, p_n \in \mathbb{N}$ , and respective utilities  $u_1, \dots, u_n \in \mathbb{R}$ , and we denote by  $P \in \mathbb{N}$  the maximal weights that we can carry. We set  $x_i = 1$  if we put the  $i$ th object in the knapsack, and  $x_i = 0$  otherwise. We want to maximize the utility of the knapsack under the weight constraint

$$\max_{x \in \{0,1\}^n} \sum_{1 \leq i \leq n} x_i u_i. \quad (11.52)$$

$$\sum_{1 \leq i \leq n} x_i p_i \leq P$$

For this, we consider, for every  $1 \leq t \leq n$  and  $0 \leq Q \leq P$ , the problem

$$\max_{\substack{x \in \{0,1\}^t \\ \sum_{1 \leq i \leq t} x_i p_i \leq Q}} \sum_{1 \leq i \leq t} x_i u_i, \quad (11.53)$$

where we denote by  $v_Q^t$  the optimal value. We denote by  $v^t = (v_Q^t)_{0 \leq Q \leq P}$ .

1. Express  $v^t$  as a function of  $v^{t-1}$  with the help of a dynamic programming equation.
2. Deduce an algorithm to solve (11.52). What is the time of execution of the algorithm?
3. Apply the algorithm to the following example:

$$\max_{\substack{x \in \{0,1\}^3 \\ 2x_1 + 3x_2 + 5x_3 \leq 6}} 8x_1 + 2x_2 + 9x_3. \quad (11.54)$$

**Exercise 11.4.6 (minimum cost path with time constraint)** Let  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  be an oriented graph, equipped with two valuations,  $c \in \mathbb{R}^{\mathcal{A}}$ , a cost, and  $\tau \in \mathbb{N}^{\mathcal{A}}$ , a time, and  $\phi \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$  a final penalty. We fix a source node  $s$  and a date limit  $T \in \mathbb{N}$ , and we look for a path  $(\ell_0, \dots, \ell_m)$  of arbitrary length, starting from  $s$  (that is,  $\ell_0 = s$ ), such that the total gain  $c_{\ell_0, \ell_1} + \dots + c_{\ell_{m-1}, \ell_m} + \phi_{\ell_m}$  is minimal, under the constraint of respecting the date limit  $T$ , that is, under the constraint  $\tau_{\ell_0, \ell_1} + \dots + \tau_{\ell_{m-1}, \ell_m} \leq T$ . We shall assume that there is a circuit whose arcs have zero time. Formulate a dynamic programming algorithm to solve this problem. Application: find by dynamic programming the minimum cost path from node 1 to node 6, in time at most 10, for the example (11.6.2) which will be treated later by Lagrangian relaxation.

**Exercise 11.4.7** Let us consider a theatre lover, who goes in July for one day to see parts of the Avignon festival. The festival has many hundreds of pieces. Each piece is characterized by a single place in the town (a theatre), and a unique time in the day (for example, 16:00-17:30). We also know the time that is necessary to go from one theatre to another. By reading the programme before going to the festival, our spectator assigns to each piece an expected pleasure, measured on a scale from 0 to 5. His aim is to see in the day a sequence of pieces, so to maximize the sum of the expected pleasures for the different pieces chosen. Show that this problem can be reduced to a minimal cost path in a graph without circuits, and that it can be solved in a time which is quadratic in the number of pieces of the festival (we will neglect time needed for meals).

**Exercise 11.4.8** We can imagine that, on the planet Mars, the extraterrestrials teach the children to count with the addition  $(a, b) \mapsto a \oplus b = \min(a, b)$ , and the multiplication  $(a, b) \mapsto a \otimes b = a + b$  (we borrow this joke from V.P. Maslov, Operational methods, MiR, Moscow, 1976). The corresponding algebraic structure,  $(\mathbb{R} \cup \{+\infty\}, \oplus, \otimes)$  is called the **minplus semiring**. It satisfies the same axioms as rings, except that addition, instead of being a group law, satisfies  $a \oplus a = a$ . For our Martians, the Bellman equation associated with the problem at finite horizon (11.41) is none other than a minplus matrix product

$$v_i^T = \bigoplus_{k \in \mathcal{N}} M_{i,k} \otimes v_k^{T-1},$$

with  $M_{i,k} = c_{i,k}$  if  $(i, k) \in \mathcal{A}$ , and  $M_{i,k} = +\infty$  otherwise. The Martians, which use the same matrix notation as us, but in the minplus semiring, write very simply,

$$v^t = M v_{t-1} \quad \text{and} \quad v^T = M^T \phi.$$

As for the Bellman equation (11.51), they write it

$$v = Mv \oplus \phi . \quad (11.55)$$

Show that the minimum cost of a path of length  $t$  going from  $i$  to  $k$  is given by  $(M^t)_{i,k}$ . Show that the maximal solution of (11.55) is equal to  $v = M^* \phi$ , where  $M^* = M^0 \oplus M \oplus M^2 \oplus \dots$ . Thus recover theorem 11.4.2. Show that if  $\mathcal{G}$  does not have a circuit of negative (or zero) cost,  $\lim_{T \rightarrow \infty} M^T = +\infty$  (the matrix all whose coefficients are equal to  $+\infty$ ). Thus recover the uniqueness result of exercise 11.4.2.

**Exercise 11.4.9** Let us say finite Markovian game for a game with two players, 'white', and 'black', who in turn move a token on a finite graph, starting from an initial position (it is white who starts). Certain vertices of the graph are final: when the token is in this vertex, we know that white has won, or black has won, or it is a draw. Can we model chess or draughts by such a game? We assume that the game always finishes. Write a dynamic programming equation expressing the value of a position for white. (Hint: this equation will use, at the same time, both min and max.) Deduce that of the three following assertions, only one is true: white can always win; black can always win; white and black can always be forced to draw.

**Remark 11.4.6** Having arrived at this point, the reader could have the impression that everything can be solved by dynamic programming. This is almost true and dynamic programming is a powerful tool, except that, just like Markovian methods in probability, dynamic programming is subject to what we call the 'curse of dimensionality': in the case of truly difficult problems, the necessary state space can be very large (think of the game of chess). •

## 11.5 Greedy algorithms

### 11.5.1 General points about greedy methods

We say that an algorithm to minimize a criterion is **greedy**, if it constructs an admissible solution by taking a sequence of decisions, where we take the best decision at each step according to a local criterion, and never reconsidering the preceding decisions. When the admissible solutions thus obtained are suboptimal, we talk of a **greedy heuristic**. For example, if we have a certain number of parcels of various sizes to put in containers, and if we want to minimize the number of containers used (this is a version of the 'packing' problem), we can imagine a method which consist of arranging the parcels as a function of their volume, and putting the parcels in the containers starting with the largest: this is a typical example of a greedy heuristic. The interest in greedy heuristics is that they are often very simple to implement. Their defect is obviously their myopia, thus the difficulty of evaluating how far their solution is from the optimum. There is however a particular class of problems for which a greedy method gives an optimal solution (see remark 11.5.5). We will be content with presenting, in the paragraph which follows, a fundamental example of a greedy algorithm giving an optimal solution.

### 11.5.2 Kruskal's algorithm for the minimum spanning tree problem

We are now interested in a nonoriented graph. Let us denote by  $\mathcal{V}$  the set of vertices and  $\mathcal{E}$  the set of edges, which is a subset of the set of the pairs of two elements of  $\mathcal{V}$ . We therefore have  $\{i, j\} \in \mathcal{E}$  if there is an edge linking the vertices  $i$  and  $j$ . Let us note that to distinguish the nonoriented case from the oriented case, we speak of vertices and edges, instead of nodes and arcs for an oriented graph.

A (nonoriented) graph is called **connected** if two arbitrary vertices can be linked by a path. An arbitrary (nonoriented) graph can be decomposed into **connected components**, which are by definition the equivalence classes for the relation  $R$  such that  $iRj$  if there is a path linking  $i$  and  $j$ . We say **forest** for a (nonoriented) graph without a circuit. A **tree** is a connected forest. We say that a subgraph  $\mathcal{G}'$  covers a graph  $\mathcal{G}$  if each vertex of  $\mathcal{G}$  is an extremity of at least one edge of  $\mathcal{G}'$ . Given a (nonoriented) connected graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , whose edges are equipped with a cost function  $\mathcal{E} \rightarrow \mathbb{R}$ ,  $\{i, j\} \mapsto c_{ij}$ , the problem of the **minimum spanning tree** consists of finding a covering tree  $\mathcal{T}$  whose cost

$$\sum_{\{i,j\} \in \mathcal{T}} c_{ij},$$

is minimum. This optimization problem is met, for example, in cabling problems, when we want to connect electrically a set of points to each other by minimizing the length of wire.

The Kruskal algorithm constructs a sequence of forests. We start from the forest comprising all the vertices and no edge. At each step, we choose to add to the forest, among all the edges whose addition does not create a circuit, that whose cost is minimum. The algorithm finishes when the forest is a covering tree.

**Theorem 11.5.1** *If  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  is a (nonoriented) connected graph, equipped with an arbitrary cost function  $c : \mathcal{E} \rightarrow \mathbb{R}$ , the Kruskal algorithm gives a minimum spanning tree.*

Before showing theorem 11.5.1, we state a very elementary property of covering trees, whose verification is left to the reader as an exercise.

**Lemma 11.5.2 (exchange lemma)** *If  $\mathcal{T}$  is a covering tree of a graph  $\mathcal{G}$ , and if  $i$  and  $j$  are two vertices of  $\mathcal{G}$ , there exists a unique path of  $\mathcal{T}$  linking  $i$  to  $j$ . Moreover, if  $\{i, j\}$  is an edge of  $\mathcal{G}$  which does not belong to  $\mathcal{T}$ , we again obtain a covering tree if we replace an arbitrary edge of the path of  $\mathcal{T}$  linking  $i$  to  $j$  by  $\{i, j\}$ .*

We can now state the optimality condition.

**Lemma 11.5.3** *Let  $\mathcal{G}$  and  $c$  be as in theorem 11.5.1, and let  $\mathcal{T}$  be a tree covering  $\mathcal{G}$ . The following assertions are equivalent.*

- (1)  $\mathcal{T}$  has minimum cost;
- (2) for each edge  $\{i, j\}$  which is not in  $\mathcal{T}$ , the unique path of  $\mathcal{T}$  linking  $i$  and  $j$  is formed of edges each with cost less than or equal to  $c_{ij}$ ;
- (3) for each edge  $\{r, s\}$  of  $\mathcal{T}$ , for every connected component  $C$  of the graph obtained by removing  $\{r, s\}$  from  $\mathcal{T}$ , and for every edge  $\{i, j\}$  with exactly one end in  $C$ ,  $c_{rs} \leq c_{ij}$ .

**Proof.** The implication (not 2) $\Rightarrow$ (not 1) follows from the exchange lemma 11.5.2: if  $\{i, j\}$  is an edge which is not in  $\mathcal{T}$ , and if  $\{r, s\}$  is an edge of the unique path of  $\mathcal{T}$  linking  $i$  and  $j$ , such that  $c_{ij} < c_{rs}$ , we obtain by using  $\{i, j\}$  in place of  $\{r, s\}$  in  $\mathcal{T}$  a new covering tree of cost strictly less than that of  $\mathcal{T}$ .

Let us show now (not 3) $\Rightarrow$ (not 2). Let us remark first of all that the graph obtained by removing  $\{r, s\}$  from  $\mathcal{T}$  has exactly two connected components,  $C$  and  $\overline{C} = \mathcal{V} \setminus C$ . Let us assume that we have an edge  $\{i, j\}$  with  $i \in C$  and  $j \in \overline{C}$ , such that  $c_{ij} < c_{rs}$ . The unique path linking  $i$  to  $j$  in  $\mathcal{T}$  must contain  $\{r, s\}$ , which shows that condition (2) is not satisfied.

Let us show finally (3) $\Rightarrow$ (1). Let  $\mathcal{T}$  be a tree satisfying (3), and let  $\mathcal{T}'$  be a tree of optimal cost, that we can choose such that the number of common edges between  $\mathcal{T}$  and  $\mathcal{T}'$  is maximal. We shall show that  $\mathcal{T} = \mathcal{T}'$ . In the opposite case, we can find an edge  $\{r, s\}$  in  $\mathcal{T}$  and not in  $\mathcal{T}'$ . The graph obtained by adding  $\{r, s\}$  to  $\mathcal{T}'$  contains a circuit,  $\mathcal{C}$ , containing the edge  $\{r, s\}$ . Let us now consider the two connected components  $C$  and  $\overline{C}$  of the graph obtained by removing  $\{r, s\}$  from  $\mathcal{T}$ . As the circuit  $\mathcal{C}$  already contains an edge linking  $C$  to  $\overline{C}$ , that is  $\{r, s\}$ , it must contain another edge,  $\{i, j\}$ , linking  $C$  and  $\overline{C}$ . The graph  $\mathcal{T}''$  obtained by replacing  $\{i, j\}$  by  $\{r, s\}$  in  $\mathcal{T}'$  is again a covering tree, from the exchange lemma, and condition (3) shows that the cost of  $\mathcal{T}''$  is less than or equal to that of  $\mathcal{T}'$ . Thus,  $\mathcal{T}''$  is again a covering tree of minimum cost, and as  $\mathcal{T}''$  has, in common with  $\mathcal{T}$ , one edge more than  $\mathcal{T}'$ , we contradict the maximality hypothesis in the definition of  $\mathcal{T}'$ . We have shown  $\mathcal{T}' = \mathcal{T}$ .  $\square$

**Proof of theorem 11.5.1.** It is immediate that the Kruskal algorithm, applied to a connected graph, finishes with a covering tree. This tree satisfies assertion 2 of lemma 11.5.3.  $\square$

**Remark 11.5.4** It is possible to implement the Kruskal algorithm in time  $O(|\mathcal{E}| \log |\mathcal{E}|)$  where  $|\mathcal{E}|$  is the number of edges of the graph, see [11].  $\bullet$

**Remark 11.5.5** The optimality of greedy methods is always due to strong properties of structure. For example, the reader has already met a greedy algorithm in linear algebra: we can view the problem consisting of constructing a basis of a finite dimensional vector space  $E$  as an optimization problem, consisting of maximizing the cardinality of a linearly independent family of  $E$ . The algorithm which starts from an empty family, and at each step, adds to the family an arbitrary vector of  $E$  which is not a linear combination of the vectors already in the family, is a greedy algorithm (which gives an optimal solution, that is,

a basis). The proof that the method works is based on a simple result of linear algebra, the exchange lemma, readers will already have noticed the analogy with lemma 11.5.2 above. More generally, the properties which allow us to show that the greedy algorithm is correct have been studied in the framework of the theory of the matroids and antimatroids, see particularly [11]. •

## 11.6 Separation and relaxation of combinatorial problems

In the solution of an OR problem, one of the first tasks is to recognize if an efficient exact method (path algorithm, flow algorithm, greedy algorithm, etc.) is applicable. What can we do, however, when even with good modelling, polynomial methods cannot be applied? We can of course revert to particular heuristics, or to metaheuristics such as simulated annealing or tabu search, see remark 11.6.7. In the spirit of this course, we concentrate on the exact methods, based on mathematical programming and tree searches, which allow us to prove the optimality of the solution found, or at least, to measure the variation from the optimum.

### 11.6.1 Separation and evaluation (branch and bound)

Let us consider the very general combinatorial problem

$$\min_{x \in X} J(x), \quad (11.56)$$

with  $X$  finite (but large), and  $J : X \rightarrow \mathbb{R}$ . In the method of separation and evaluation (or commonly ‘branch and bound’), the **separation** consists of representing the set  $X$  of the admissible solutions by the leaves of a tree, which we will explore. The internal nodes of the tree represent the partial decisions (corresponding to fixing some decision variables, but not all), the node at the root represents an initial situation, at which we have made no decisions. The **evaluation** is concerned, for each internal node  $s$  of the tree, with the **conditional cost**

$$v(s) = \min_{s' \text{ is a leaf which is descendant from } s} J(s'). \quad (11.57)$$

(The word ‘descendant’ has the genealogical meaning, that is, we orient the tree with the root at the top, and the leaves at the bottom.) The calculation of this conditional cost at  $s$  is often as hard as that of the initial problem (11.56) (in particular, when  $s$  is the root, (11.57) coincides with (11.56)), this is why we shall simplify the problem (11.57), by allowing ourselves the introduction of a lower bound  $b(s)$  of the conditional cost at  $s$ :

$$b(s) \leq v(s), \quad (11.58)$$

that we will have to define for every internal node of the tree.



### Statement of the separation and evaluation algorithm

The separation and evaluation algorithm consists of exploring the nodes of the tree, starting from the root. During the exploration, we remember  $m$ , the minimal cost of the solutions already found, therefore of the corresponding solution. We initialize the algorithm with  $m = +\infty$ . When we visit an internal node  $s$  of the graph for the first time, we evaluate the bound  $b(s)$ . If  $b(s) \geq m$ , we do not explore the daughter branches of the node  $s$ , since the cost of the leaves there is not better than the cost of the best solution already found, and we go back up to the father node of  $s$  in order to continue the exploration of the tree (we can visualize this by saying that we **cut** the branch of the tree from the node  $s$ ). If conversely  $b(s) < m$ , it is possible that the branch starting from  $s$  contains a better solution  $m$ : in this case we continue the exploration of the tree, going to the daughter nodes of  $s$ . When we arrive at a leaf of the tree which represents an admissible solution  $x \in X$  of (11.56), it only remains to calculate the value  $J(x)$ : if  $J(x) < m$ , the best solution found is  $x$ , we therefore set  $m = J(x)$ , and we remember  $x$  instead of the old best solution. We then continue the exploration of the tree by returning to the father node of the current leaf.

The algorithm visits each leaf at most once. The worst case is that where the bound  $b$  never allows us to cut a branch: the algorithm reduces in this case to listing all the admissible solutions of (11.56).

### Illustration: example of the travelling salesman

Before detailing the method, let us consider the travelling salesman problem, already stated in example 9.1.4. We shall here treat the nonoriented version of the travelling salesman problem, which considers a **complete** nonoriented graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  (complete means that  $\mathcal{E}$  is formed of all the pairs of vertices of  $\mathcal{V}$ ). We will take  $\mathcal{V} = \{1, \dots, n\}$ , with  $n = |\mathcal{V}|$ . We associate with each pair of vertices  $\{i, j\}$  a time that we denote by  $t_{ij}$  (which is an abbreviation for  $t_{\{i, j\}}$ , we therefore have  $t_{ij} = t_{ji}$  since the graph is nonoriented). The aim is to find a circuit, that is, a sequence of vertices  $\ell_1, \dots, \ell_n$  comprising each vertex of  $\mathcal{G}$  once and only once, and such that the total time

$$t_{\ell_1 \ell_2} + t_{\ell_2 \ell_3} + \dots + t_{\ell_n \ell_1} \quad (11.59)$$

is a minimum.

By using the invariance of the criterion (11.59) by cyclic permutation, we can always suppose that we start from the vertex 1, or  $\ell_1 = 1$ . The circuit is then specified uniquely by the sequence  $\ell_2, \dots, \ell_{n-1}$ . The **separation** of the problem reduces to organizing the choice of a circuit in a sequence of decisions. For example, we can consider the choice of  $\ell_2 \in \mathcal{V} \setminus \{\ell_1\}$  as a first decision, the choice of  $\ell_3 \in \mathcal{V} \setminus \{\ell_1, \ell_2\}$  as a second decision, and so on up to  $\ell_{n-1}$ . An internal vertex of the tree of separation will therefore correspond to a subsequence  $(1 = \ell_1, \ell_2, \dots, \ell_k)$ , with  $k \leq n - 2$ , which we shall call a ‘partial circuit’. It will bound below the conditional cost (11.57), that is, bound the total time of the circuits which start with  $\ell_1, \dots, \ell_k$ . We can give, for

example, the naive bound

$$b_1(\ell_1, \dots, \ell_k) = t_{\ell_1 \ell_2} + \dots + t_{\ell_{k-1} \ell_k} + \min_{j \in \mathcal{V} \setminus \{\ell_1, \dots, \ell_{k-1}\}} t_{\ell_k j} + \min_{m \in \mathcal{V} \setminus \{\ell_2, \dots, \ell_k\}} t_{m \ell_1} + (n - k - 2) \left( \min_{\substack{j, m \in \mathcal{V} \setminus \{\ell_1, \dots, \ell_k\} \\ j \neq m}} t_{jm} \right). \quad (11.60)$$

In effect, the time of the circuit (11.59) is the sum of the time of the partial circuit  $(\ell_1, \ell_2, \dots, \ell_k)$ , that is  $t_{\ell_1 \ell_2} + \dots + t_{\ell_{k-1} \ell_k}$ , plus the time of the edge  $\{\ell_k, \ell_{k+1}\}$ , that we have bounded below by the first min in (11.60), plus the time of the edge  $\{\ell_n, \ell_1\}$ , that we bound symmetrically by the second min in (11.60), and finally, the time of the path  $(\ell_{k+1}, \dots, \ell_n)$ . As this path is of length  $n - k - 2$  and does not contain any vertex of  $\{\ell_1, \dots, \ell_k\}$ , we can bound its time below by the last min in (11.60), which shows that  $b_1(\ell_1, \dots, \ell_k) \leq v(\ell_1, \dots, \ell_k)$ .

Let us now apply the separation and evaluation algorithm, with the bound  $b_1$ , to a small example of a travelling salesman. Let us consider the town to be like Manhattan represented in Figure 11.3. We shall take the set  $\mathcal{V} = \{1, \dots, 5\}$  whose elements correspondent to the points represented on the picture, with coordinates  $P_1 = (0, 0)$ ,  $P_2 = (3, 0)$ ,  $P_3 = (1, 1)$ ,  $P_4 = (3, 2)$ , and  $P_5 = (0, 3)$  (the frame of reference is towards the East and the South), and  $t_{ij}$  will represent the journey time from the point  $P_i$  to the point  $P_j$ , that is, the norm  $\|P_i - P_j\|_1$ .

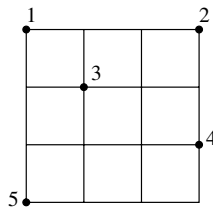


Figure 11.3. A travelling salesman in Manhattan.

The course through the corresponding tree is represented on Figure 11.4. We start from the node (1), which corresponds to an empty partial circuit starting from the point 1 of the town. The first step of the algorithm consists of choosing the following point of the town that we shall visit, which can be 2, 3, 4, or 5. Let us choose, for example, the point 2, which brings us to node (1, 2) that we have represented on the left of the tree, with the partial circuit to which it corresponds. As we have no admissible solutions for the moment,  $m = +\infty$ , and as the test  $b_1(1, 2) < m$  is automatically satisfied, we do not calculate  $b_1(1, 2)$ . Let us follow the path deeper: we arrive at node (1, 2, 3) (we have again represented the partial circuit), then to node (1, 2, 3, 4), which is a leaf, since it uniquely determines the admissible solutions  $x = (1, 2, 3, 4, 5)$ , with time  $m = 16$ . The complete circuit thus obtained is represented at the leaf. Having arrived at a leaf, the search returns to the father node, to go back

down to the following leaf, which represents the solution  $(1, 2, 3, 5, 4)$ , with time 18 worse than  $m$ . The next new node explored is  $(1, 2, 4)$ , and we calculate  $b_1(1, 2, 4) = 13 < m$ : we therefore explore the first daughter of  $(1, 2, 4)$ , which gives the solution  $x = (1, 2, 4, 5, 3)$ , with time 14, which is better than  $m$ : we therefore set  $m = 14$ . The other daughter of  $(1, 2, 4)$  gives another solution with time 14,  $(1, 2, 4, 3, 5)$ . The next new node visited is  $(1, 2, 5)$ , with  $b_1(1, 2, 5) = 17 \geq m$ : we therefore cut the subtree starting from  $(1, 2, 5)$ , and the next new node is  $(1, 3)$ . We leave the reader to finish the calculation, and to show thus that  $x = (1, 2, 4, 5, 3)$  is a circuit whose time 14 is optimal.

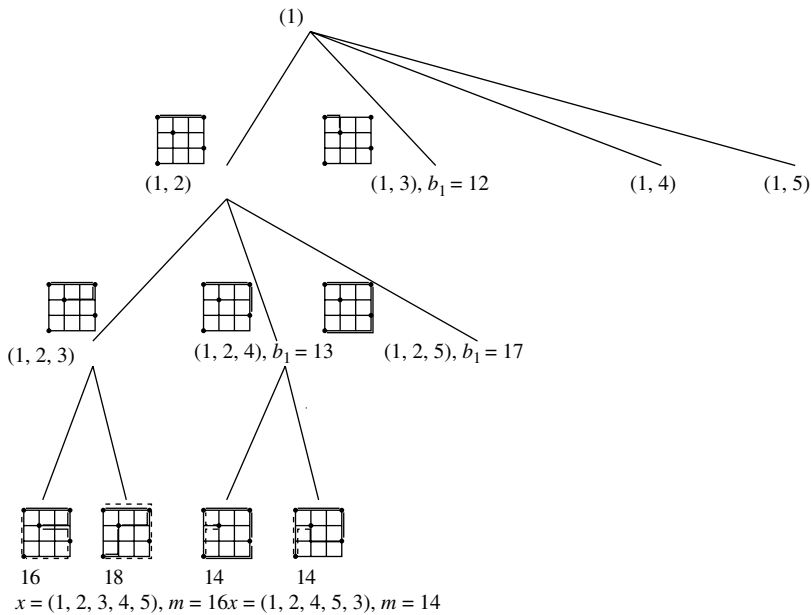


Figure 11.4. Part of the separation and evaluation tree, for the travelling salesman problem of Figure 11.3.

**Remark 11.6.1** The travelling salesman problem in Manhattan is a particular case of the travelling salesman metric, in which  $t_{ij}$  is the distance from  $i$  to  $j$  for a certain metric, here, that of the norm  $\|\cdot\|_1$ . A classical application of the travelling salesman metric is the manufacturing of printed circuits, if we consider the case of a machine having to drill a sequence of holes (according to the type of tool, we obtain a travelling salesman problem for the Euclidean norm, for the norm  $\|\cdot\|_\infty$ , etc.). The travelling salesman problem is a classical example of an NP-difficult problem (see remark 11.6.6).

### The importance of the quality of the bound

The reader may believe that the bound  $b_1$  is reasonable, but let us see how the algorithm behaves when we increase the number of vertices of the graph. We have programmed the above algorithm, and solved similar instances of the travelling salesman in Manhattan, but by varying the number  $n$  of vertices. Here is a typical set of results, giving the number of nodes of the tree of separation and evaluation, as a function of  $n$ :

$n$	5	6	7	8	9	10	11	12	13	14	15
tree	10	20	51	805	2175	10 598	58 414	199 276	499 887	1 250 530	3 598 585

The case  $n = 15$  already takes a minute on a normal PC. We have just seen what we call a **combinatorial explosion**.

Retrospectively, the coarse character of the bound  $b_1$  appears: when  $n - k$  is large, it is bad to bound the length of the complementary partial circuit of  $(\ell_1, \dots, \ell_k)$  by  $(n - k - 2)$  time the minimum of the time of the edges between the remaining vertices, this reduces to multiplying an error by a term of order  $n$ . We see here that in a separation and evaluation algorithm, it is essential to have a bound which sufficiently captures the ‘physics’ of the problem.

### The 1-tree bound for the travelling salesman

Let us give a first example of such a bound for the nonoriented travelling salesman. Given a partial circuit  $(\ell_1, \ell_2, \dots, \ell_k)$  of  $\mathcal{G}$ , with  $k \leq n - 2$ , let us construct the graph  $\mathcal{G}'$  induced by the subset of vertices  $\mathcal{V}' = \mathcal{V} \setminus \{\ell_1, \dots, \ell_k\}$ , that is, the graph of the set of vertices  $\mathcal{V}'$  and of the set of edges  $\mathcal{E}' = \{\{i, j\} \in \mathcal{E} \mid i, j \in \mathcal{V}'\}$ . Let us denote by  $\text{ac}(\mathcal{G}')$  the minimum cost of a covering tree  $\mathcal{G}'$ . (Let us recall that the idea of covering tree has been defined in Section 11.5, where we have also seen that a covering tree of minimum cost is calculated in time  $O(|\mathcal{E}| \log |\mathcal{E}|)$  by the Kruskal algorithm.) We have the following lower bound

$$\begin{aligned}
 b_2(\ell_1, \dots, \ell_k) &= t_{\ell_1 \ell_2} + \dots + t_{\ell_{k-1} \ell_k} + \min_{j \in \mathcal{V} \setminus \{\ell_1, \dots, \ell_{k-1}\}} t_{\ell_k j} \\
 &\quad + \min_{m \in \mathcal{V} \setminus \{\ell_2, \dots, \ell_k\}} t_{m \ell_1} + \text{ac}(\mathcal{G}').
 \end{aligned}
 \tag{11.61}$$

In the special case where the partial circuit is of length zero, let  $k = 1$ , we can refine the bound  $b_2$  very slightly by noting that in this case, the edges  $\{\ell_1, \ell_2\}$  and  $\{\ell_k, \ell_1\}$  of the circuit must be distinct (assuming that the graph has at least three vertices), which gives the new bound

$$b'_2(\ell_1) = \min_{j, m \in \mathcal{V} \setminus \{\ell_1\}, j \neq m} (t_{\ell_1 j} + t_{m \ell_1}) + \text{ac}(\mathcal{G}').
 \tag{11.62}$$

This last, classical, bound is known under the name of **1-tree bound** (a 1-tree of a graph is a subgraph formed on the one hand by a tree covering all vertices except a distinct vertex noted ‘1’, and on the other hand by two distinct edges where 1 is the extremity). For example, for the graph of Figure 11.3, the Kruskal algorithm

gives the 1-tree of minimal cost represented on Figure 11.5. The cost of this 1-tree is  $b'_2(1) = 13$ . Returning to the tree of Figure 11.4, we see that to replace  $b_1$  by  $b_2$  would have allowed us not to visit the descendants of node  $(1, 3)$ , as  $b_2(1, 3) = 14$  is greater (in fact, equal) to the value  $m = 14$  of the best solution met before at the visit of  $(1, 3)$ .

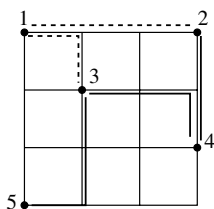


Figure 11.5. The 1-tree bound  $b'_2(1)$ , for the travelling salesman problem of Figure 11.3. The covering tree  $\mathcal{G}'$  is in thick lines, the two edges connecting this tree to the vertex 1 are in dotted lines.

### Choice of the tree and the order of exploration of the branches

There are in general many ways to represent the set of admissible solutions  $X$ , and the choice is often suggested by the technique being used to construct the bound. Thus, we see in Section 11.6.2 that we can model the travelling salesman by an integer linear programming problem, by introducing for each  $\{i, j\} \in \mathcal{E}$  a variable  $x_{ij}$  being 1 if the edge  $\{i, j\}$  belongs to the circuit considered and 0 otherwise. With such a model, we can construct a binary separation and evaluation tree, in which an elementary decision consists of fixing the value of a variable  $x_{ij}$  at 0 or 1. Another determining parameter is the order in which we visit the vertices: it is judicious to examine as soon as possible (that is, close to the root of the tree) the decisions that we think have the most influence on the cost of the solution, the aim being to cut the branches as high as possible in the tree. In the same spirit, initializing  $m$  with the value of a solution obtained heuristically, instead of  $+\infty$ , only helps to cut the branches earlier.

### 11.6.2 Relaxation of combinatorial problems

A systematic way to obtain lower bounds for the optimal value of the cost of a combinatorial problem consists of **relaxing** the problem, that is, to make the admissible set bigger in such a way as to obtain an easier problem, giving a lower bound for the initial problem. When the admissible set is represented by the constraints, a way to relax is very simply to ignore some constraints. We have already seen an example of relaxation with the bound of the 1-tree for the travelling salesman: the circuits are exactly the 1-trees such that each vertex has two neighbours. We shall now present some general techniques of relaxation.

### Continuous relaxation

The efficiency of the linear programming tools often suggest modelling combinatorial problems by integer linear programming problems: we then obtain a lower bound by relaxing the integer constraints.

To illustrate this idea let us present the relaxation of the travelling salesman problem proposed by Dantzig, Fulkerson, and Johnson in an article in 1954, which at that time solved a problem with 49 towns. This was the starting point of a series of works allowing us today to solve exactly problems with many thousands of towns (for a history and a recent state of the art, we refer to D. Applegate, R. Bixby, V. Chvátal, and W. Cook, ‘On the solution of traveling salesman problems’, *Documenta Math.*, Extra volume ICM 1998, III, 645–656, and more generally on the web <http://www.math.princeton.edu/tsp>).

We use the notation of the preceding section for the travelling salesman problem:  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  is a complete nonoriented graph (that is,  $\mathcal{E}$  contains all the edges linking two elements of  $\mathcal{V}$ ), with a function time  $t : \mathcal{E} \rightarrow \mathbb{R}$ ,  $\{i, j\} \mapsto t_{ij}$ . With each circuit, we associate a vector  $x \in \{0, 1\}^{\mathcal{E}}$  such that  $x_{ij} = 1$  if  $\{i, j\}$  is part of the circuit, and  $x_{ij} = 0$  otherwise. Conversely, a vector  $x \in \{0, 1\}^{\mathcal{E}}$  represents the subgraph of  $\mathcal{G}$  having for edges the  $\{i, j\}$  such that  $x_{ij} = 1$ . We must now express by the linear constraints the fact that  $x \in \{0, 1\}^{\mathcal{E}}$  represents a circuit. As each vertex of a circuit has exactly two neighbours,  $x$  must satisfy

$$\sum_{k \in \mathcal{V}, \{k, j\} \in \mathcal{E}} x_{kj} = 2, \quad \text{for every } j \in \mathcal{V}. \quad (11.63)$$

The constraints (11.63) are not sufficient to characterize a circuit, since the set of edges  $\{i, j\}$  such that  $x_{ij} = 1$  can be disconnected: in fact, we can see that the  $x \in \{0, 1\}^{\mathcal{E}}$ , the solutions of (11.63), exactly represent the disjoint unions of circuits. In order to eliminate the parasitical solutions, we can add the following constraints, called subtour inequalities,

$$\text{for every } \mathcal{S} \subset \mathcal{V}, \text{ such that } \mathcal{S} \neq \emptyset \text{ and } \mathcal{S} \neq \mathcal{V}, \quad \sum_{k, m \in \mathcal{S}, \{k, m\} \in \mathcal{E}} x_{km} \leq |\mathcal{S}| - 1. \quad (11.64)$$

We easily see that  $x \in \{0, 1\}^{\mathcal{E}}$  satisfies (11.63) and (11.64) if, and only if, it represents a circuit: in effect, every vector associated with a circuit satisfies these constraints, and conversely, as every  $x \in \{0, 1\}^{\mathcal{E}}$  satisfying (11.63) represents a disjoint union of circuits, it is enough to take for  $\mathcal{S}$  the set of vertices of an arbitrary circuit to obtain  $\sum_{k, m \in \mathcal{S}, \{k, m\} \in \mathcal{E}} x_{km} = |\mathcal{S}|$ , and if  $x$  satisfies the subcircuit inequalities (11.64), we must have  $\mathcal{S} = \mathcal{V}$ , which shows that  $x$  represents a circuit. This allows us to formulate the travelling salesman problem as an integer linear programming problem.

$$(VC) : \quad \min \sum_{\{i, j\} \in \mathcal{E}} t_{ij} x_{ij} \quad \text{under the constraints } x \in \{0, 1\}^{\mathcal{E}}, \quad (11.63), (11.64),$$

and we immediately obtain a lower bound by considering the relaxed linear programming problem in continuous variables

$$(VC)_{\text{rel}} : \min \sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij} \quad \text{under the constraints } 0 \leq x_{ij} \leq 1, \quad (11.63), (11.64).$$

Despite the exponential number of constraints in (11.64), it is possible to solve  $(VC)_{\text{rel}}$  efficiently by proceeding as follows. We start by minimizing  $\sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij}$  under the constraints  $0 \leq x_{ij} \leq 1$  and (11.63), ignoring the subcircuit constraints (11.64). We thus find a first  $x \in [0, 1]^{\mathcal{E}}$ . We then see if there exists a subset  $\mathcal{S} \subset \mathcal{V}$ ,  $\mathcal{S} \neq \emptyset$ ,  $\mathcal{S} \neq \mathcal{V}$ , for which the inequality of (11.64) is violated, which can be done (exercise 11.6.1) very efficiently. If none exists, we have solved  $(VC)_{\text{rel}}$ . If conversely we have found  $\mathcal{S}$  such that (11.64) is not satisfied, we again minimize  $\sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij}$  under the constraints  $0 \leq t_{ij} \leq 1$  and (11.63), by adding the subcircuit inequality associated with  $\mathcal{S}$ . Following this sequence of minimizations and successive additions of inequalities, we are finally led to a solution of  $(VC)_{\text{rel}}$ . This method can be interpreted geometrically by speaking of **intersections**. Indeed let us denote by  $P$  the polytope of the admissible solutions of problem  $(VC)_{\text{rel}}$ , and let  $t : x \mapsto \sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij}$  be the linear form that we minimize. The method that we have outlined reduces to defining a decreasing sequence of polytopes  $P^1 \supset P^2 \supset \dots \supset P$ . We take, first of all, for  $P^1$  the set defined by  $0 \leq x_{ij} \leq 1$  and (11.63). We minimize  $t$  over  $P^1$ , and the minimum is attained at a point  $x^1$ . If  $x^1$  is in  $P$ , we have solved  $(VC)_{\text{rel}}$ , otherwise detecting a constraint (11.64) violated by  $x^1$  reduces to **separating**  $x^1$  from  $P$ , that is of finding a particular half-space  $H^1$  such that  $x^1 \notin H^1$ , and  $P \subset H^1$ , and the following step reduces to minimizing  $t$  on the new polytope  $P^2 = P^1 \cap H^1$ , obtained by ‘intersecting’  $P^1$  by  $H$ . We therefore construct a decreasing sequence of polytopes of which we can say, intuitively, that they ‘approximate’  $P$  in the neighbourhood of the point where  $t$  is minimal. Such techniques can even give the solution of the original problem  $(VC)$ , if we can find other intersections, of a combinatorial nature, allowing us to approximate sufficiently well the integer closure  $P_e$  of  $P$ .

**Exercise 11.6.1** Show that in the problem  $(VC)_{\text{rel}}$ , we can replace the constraints (11.64) by the constraints

$$\text{for every } \mathcal{S} \subset \mathcal{V}, \text{ such that } \mathcal{S} \neq \emptyset \text{ and } \mathcal{S} \neq \mathcal{V}, \quad \sum_{k \in \mathcal{V}, m \in \mathcal{S} \setminus \mathcal{V}, \{k,m\} \in \mathcal{E}} x_{km} \geq 2. \quad (11.65)$$

Show, by using the Ford–Fulkerson theorem (Question 4 of exercise 11.3.8) that we can check if a vector  $x \in [0, 1]^{\mathcal{E}}$  satisfies (11.65) with the help of a flow algorithm.

## Lagrangian relaxation

Let us consider the very general problem

$$\begin{aligned} &\text{minimize } J(x) \text{ under the constraints:} \\ &\quad x \in X, \\ &\quad F_i(x) \leq 0, \quad i = 1, \dots, m, \\ &\quad F_i(x) = 0, \quad i = m+1, \dots, m+q, \end{aligned} \tag{11.66}$$

where  $J, F_1, \dots, F_{m+q}$  are functions from  $\mathbb{R}^n$  into  $\mathbb{R}$ , and  $X$  is a nonempty subset of  $\mathbb{R}^n$ . Let us consider the Lagrangian  $\mathcal{L} : X \times \Lambda \rightarrow \mathbb{R}$ , with  $\Lambda = (\mathbb{R}_+)^m \times \mathbb{R}^q$  and

$$\mathcal{L}(x, \lambda) = J(x) + \lambda_1 F_1(x) + \dots + \lambda_{m+q} F_{m+q}(x).$$

Let  $J^*$  be the optimal value of (11.66). We have already noted in remark 10.3.10 that we always have the weak duality inequality

$$J^* = \inf_{x \in X} \sup_{\lambda \in \Lambda} \mathcal{L}(x, \lambda) \geq \sup_{\lambda \in \Lambda} \inf_{x \in X} \mathcal{L}(x, \lambda) = \sup_{\lambda \in \Lambda} \mathcal{D}(\lambda), \tag{11.67}$$

where

$$\mathcal{D} : \Lambda \rightarrow \mathbb{R} \cup \{-\infty\}; \quad \mathcal{D}(\lambda) = \inf_{x \in X} \mathcal{L}(x, \lambda) \tag{11.68}$$

is the dual function. The method of **Lagrangian relaxation** consists of using the right-hand side of (11.67) as a lower bound for the value  $J^*$  of the original problem (11.66). Every Lagrange multiplier  $\lambda \in \Lambda$  gives a bound  $\mathcal{D}(\lambda) \leq J^*$ , but it is natural to look for the best bound possible, which reduces to maximizing  $\mathcal{D}$ . Now  $\mathcal{D}$ , which is an infimum of affine functions, is concave. If, as is the case for most of the combinatorial problems,  $X$  is finite,  $\mathcal{D}$  which is a finite infimum of affine functions, is a **nondifferentiable** function (except of course in the degenerate cases). Maximizing  $\mathcal{D}$  therefore raises the problem of **nondifferentiable optimization**, which treats the minimization of nondifferentiable convex functions (or maximization of nondifferentiable concave functions).

We shall briefly present a very simple method, the **subgradient method**, which generalizes the gradient methods of Section 10.5.2. We must first define the ideas of **subgradient** (for convex functions) or of **supergradient** (for concave functions) which are fundamental ideas in convex analysis (we will be content with a short review). It will be useful to consider convex functions with values in  $\mathbb{R} \cup \{+\infty\}$ , and, symmetrically, concave functions with values in  $\mathbb{R} \cup \{-\infty\}$ , since for example the function  $\mathcal{D}$  defined by (11.68) is concave and can take the value  $-\infty$  (the value  $+\infty$  is not possible, as we have excluded the case  $X = \emptyset$ ). When  $J$  is convex from  $\mathbb{R}^n$  to  $\mathbb{R} \cup \{+\infty\}$ , let us denote by  $\text{dom } J = \{x \in \mathbb{R}^n \mid J(x) < +\infty\}$  the **domain** of  $J$ . Symmetrically, if  $J$  is a concave function from  $\mathbb{R}^n$  into  $\mathbb{R} \cup \{-\infty\}$ , we define  $\text{dom } J = \{x \in \mathbb{R}^n \mid J(x) > -\infty\}$ . It is useful only to treat the case of convex functions, given that all the results have symmetrical versions for concave functions. The reader can symmetrize the results to be applied to the maximization of dual functions associated with minimization problems.



**Definition 11.6.2** If  $J$  is a convex function of  $\mathbb{R}^n$  in  $\mathbb{R} \cup \{+\infty\}$ , and if  $x \in \text{dom } J$ , we say **subdifferential** of  $J$  at  $x$  for the set

$$\partial J(x) = \{p \in \mathbb{R}^n \mid J(y) - J(x) \geq p \cdot (y - x), \quad \forall y \in \mathbb{R}^n\}. \quad (11.69)$$

The elements of  $\partial J(x)$  are called **subgradients**. The ideas of superdifferential and supergradient of a concave function of  $\mathbb{R}^n$  in  $\mathbb{R} \cup \{-\infty\}$  are defined symmetrically by reversing the inequality in (11.69).

It follows immediately from definition (11.69) that the subdifferential  $\partial J(x)$  is a convex closed set of  $\mathbb{R}^n$ . On the other hand, if  $0 \in \partial J(x)$ ,  $x$  is obviously a minimum point of  $J$ .

**Remark 11.6.3** The subdifferentials can be defined by (11.69) even when  $J$  is not convex. They are, however, especially useful when  $J$  is convex, as in this case  $\partial J(x) \neq \emptyset$  at every point  $x$  in the interior of  $\text{dom } J$  (this is a consequence of the geometrical form of the Hahn–Banach theorem). The convexity also guarantees the equivalence of definition 11.6.2 of the subgradients with that of the local definitions. We can in effect define the subgradients of an arbitrary function  $J$  of  $\mathbb{R}^n$  in  $\mathbb{R} \cup \{+\infty\}$ , by saying that  $p$  is a subgradient of  $J$  at a point  $x \in \text{dom } J$  if  $J$  is locally ‘above’ the affine function  $y \mapsto J(x) + p \cdot (y - x)$ , that is,

$$J(y) \geq J(x) + p \cdot (y - x) + o(y - x), \quad \text{when } y \rightarrow x.$$

When  $J$  is convex, this inequality is equivalent to  $p \in \partial J(x)$ . •

Let us now apply these ideas to the dual function  $\mathcal{D}$ . It will be convenient to extend  $\mathcal{D}$  to  $\mathbb{R}^{m+q}$  by setting  $\mathcal{D}(\lambda) = -\infty$ , if  $\lambda \notin \Lambda$ , which defines a concave function from  $\mathbb{R}^{m+q}$  into  $\mathbb{R} \cup \{-\infty\}$ . When  $X$  is finite,  $\text{dom } \mathcal{D} = \Lambda$ . A supergradient of the dual function  $\mathcal{D}$  is calculated easily, with the help of the following observation.

**Proposition 11.6.4** Let us assume that  $X$  is finite, let  $\mathcal{D}$  be the dual function defined by (11.68), and for every  $\lambda \in \Lambda$ , let us set  $\Gamma(\lambda) = \arg \max_{x \in X} \mathcal{L}(x, \lambda) = \{x \in X \mid \mathcal{L}(x, \lambda) = \mathcal{D}(\lambda)\}$ . Then, for every  $x \in \Gamma(\lambda)$ ,  $F(x) = (F_i(x))_{1 \leq i \leq m+q}$  is a supergradient of  $\mathcal{D}$  at the point  $\lambda$ .

**Proof.** For every  $x \in \Gamma(\lambda)$ , and for every  $\mu \in \Lambda$ , we have  $\mathcal{D}(\mu) - \mathcal{D}(\lambda) \geq \mathcal{L}(x, \mu) - \mathcal{L}(x, \lambda) = F(x) \cdot (\mu - \lambda)$ , which shows that  $F(x)$  is a supergradient of  $\mathcal{D}$  at the point  $\lambda$ . □

**Remark 11.6.5** Proposition 11.6.4 is a weak version of the following result on the subdifferentials of the maxima of convex functions. If  $J$  from  $\mathbb{R}^n$  into  $\mathbb{R} \cup \{+\infty\}$  is of the form  $J(x) = \sup_{i \in I} J_i(x)$ , where  $I$  is a finite set, and if  $x \mapsto J_i(x)$  is convex, for every  $i \in I$ , then, for every  $x \in \text{dom } J$ ,

$$\partial J(x) = \text{co} \left( \bigcup_{\substack{i \in I \\ J_i(x) = J(x)}} \partial J_i(x) \right). \quad (11.70)$$

The proof of this identity is the object of exercise 11.6.5. •

Let  $P_\Lambda$  be the projection from  $\mathbb{R}^{m+q}$  into  $\Lambda$ ,  $\lambda \mapsto (\lambda_1^+, \dots, \lambda_m^+, \lambda_{m+1}, \dots, \lambda_{m+q})$ . The (projected) **supergradient algorithm** to maximize the concave function  $\mathcal{D}$  consists of constructing the sequence

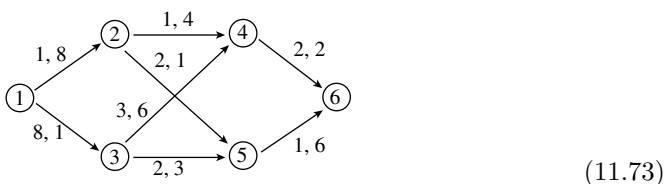
$$\lambda_{k+1} = P_\Lambda \left( \lambda_k + \frac{\rho_k}{\|p_k\|} p_k \right), \quad (11.71)$$

where  $\lambda_0 \in \Lambda$  is chosen arbitrarily, where  $p_k$  is an arbitrary supergradient of  $\mathcal{D}$  at the point  $\lambda_k$ , and where  $\rho_k$  is a sequence of strictly positive real numbers such that

$$\rho_k \rightarrow 0, \quad \sum_i \rho_i = +\infty. \quad (11.72)$$

Obviously, the value  $\lambda_{k+1}$  is only well defined if  $p_k \neq 0$ . When  $p_k = 0$ , the algorithm stops:  $\lambda_k$  is then the maximum of  $J$  (by definition of the supergradients the nullity of a supergradient at a point implies that the function is maximal in this point).

Let us now illustrate Lagrangian relaxation by treating the example of the minimum cost path with time constraint, already mentioned in exercise 11.4.6 as an application of dynamic programming. Let us therefore consider an oriented graph  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ , the arc  $(i, j)$  being equipped with the cost  $c_{ij}$  and with the time  $\tau_{ij}$ . To fix ideas, we will look for the minimum cost path and with total time at most 10, going from the source node  $s = 1$  to the sink node  $p = 6$ , in the graph



We have represented on each arc the values of  $c$  and  $\tau$ , in this order, for example, the arc  $(1, 2)$  costs  $c_{12} = 1$  and takes  $\tau_{12} = 8$  units of time.

We must first of all formulate this problem in the form (11.66). For this, we represent a path by the Boolean vector  $x \in \mathbb{R}^{\mathcal{A}}$  such that  $x_{ij} = 1$  if  $(i, j)$  belongs to the path, and  $x_{ij} = 0$  otherwise. The problem of the shortest path from a source  $s$  to a sink  $p$  in time at most  $T$  is then written

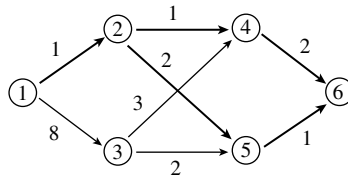
$$\begin{aligned} & \text{minimize} && \sum_{(i,j) \in \mathcal{A}} c_{ij} x_{ij} \\ & \text{under the constraints:} && x \in \{0, 1\}^{\mathcal{A}}, \\ & && x \text{ represents a path from } s \text{ to } p, \\ & && \sum_{(i,j) \in \mathcal{A}} \tau_{ij} x_{ij} \leq T. \end{aligned} \quad (11.74)$$

There are of course many ways to write a program (11.66). Thus, it is voluntarily that we have left the constraint ' $x$  represents a path from  $s$  to  $p$ ' in a literal form. We

could have clarified this constraint by using the Kirchoff law (11.31). But dualizing this law is a bad idea for the success of Lagrangian relaxation depends precisely on the capacity to identify the ‘smallest’ set of constraints whose relaxation leads to a simpler problem, preferably solvable by a direct combinatorial method. Here, it is the time constraint that we must relax, since if we ignore this constraint, we obtain a problem purely of minimum cost path. By dualizing the time constraint, the function  $\mathcal{D} : \mathbb{R}_+ \rightarrow \mathbb{R} \cup \{-\infty\}$  is written

$$\begin{aligned} \mathcal{D}(\lambda) &= \inf_{\substack{x \in \{0,1\}^{\mathcal{A}} \\ x \text{ represents a path from } s \text{ to } p}} \left( \sum_{(i,j) \in \mathcal{A}} c_{ij} x_{ij} \right) + \lambda \left( \sum_{(i,j) \in \mathcal{A}} \tau_{ij} x_{ij} - T \right) , \\ &= -\lambda T + \inf_{\substack{x \in \{0,1\}^{\mathcal{A}} \\ x \text{ represents a path from } s \text{ to } p}} \sum_{(i,j) \in \mathcal{A}} (c_{ij} + \lambda \tau_{ij}) x_{ij} . \end{aligned}$$

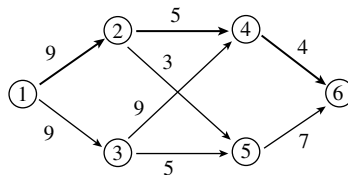
With fixed  $\lambda$ , the calculation of  $\mathcal{D}(\lambda)$  reduces to solving a classical problem of minimum cost path **without time constraints** in the graph of costs  $c_{ij} + \lambda \tau_{ij}$ , this which can be done by dynamic programming, and which takes a linear time in the case of a graph without circuits, as is the case, for example, (11.6.2). Let us now apply the supergradient algorithm (11.71) to this example. For  $\lambda_1 = 0$ , we must find the minimum cost path  $1 \rightarrow 6$  in the graph equipped with the costs  $c_{ij} + 0\tau_{ij}$



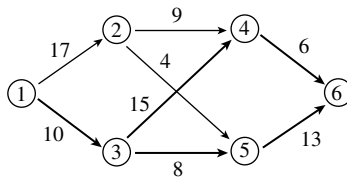
The set  $\Gamma(0)$  of the optimal paths is reduced to the paths  $(1, 2, 4, 6)$ , and  $(1, 2, 5, 6)$  represented in thick lines, which have cost 4. Thus,  $\mathcal{D}(0) = 4 - 0T = 4$ . From proposition 11.6.4, we have the supergradient

$$p_1 = \tau_{12} + \tau_{24} + \tau_{46} - T = 4 .$$

With a step  $\rho_1 = 1$ , (11.71) gives  $\lambda_2 = 1$ , and the new graph equipped with the costs  $c_{ij} + 1\tau_{ij}$



As the optimal path has not changed we have the same supergradient  $p_2 = p_1$ . By taking, for example,  $\rho_2 = 1$ , we have  $\lambda_3 = 2$ , which gives the graph



There are this time two optimal paths:  $\Gamma(\lambda_3) = \{(1, 3, 4, 6), (1, 3, 5, 6)\}$ , of cost 21. In particular, the supergradient corresponding to the path  $(1, 3, 5, 6)$  is  $\tau_{13} + \tau_{35} + \tau_{56} - T = 0$ . The supergradient algorithm therefore stops:  $\max_{\lambda \in \mathbb{R}_+} \mathcal{D}(\lambda) = \mathcal{D}(2) = 31 - 20 = 11$ . As the path  $(1, 3, 5, 6)$  has time  $10 \leq T$  and cost  $11 = \mathcal{D}(2)$ , we solve not only the relaxed Lagrangian problem, which consists of maximizing  $\mathcal{D}(\lambda)$ , but also the initial problem (11.74). The fact that the relaxed Lagrangian provides a solution of the initial problem is, however, exceptional. In general, there is a duality gap, but the  $x$  achieving the min in the dual function (11.68), evaluated at a maximum point  $\lambda$  of  $\mathcal{D}$ , can often be modified without adding too much to the cost to arrive at a (suboptimal) solution of the initial problem. We talk then of **Lagrangian heuristics**.

To conclude this illustration of Lagrangian relaxation, let us revisit the travelling salesman problem in a complete nonoriented graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , equipped with a time function  $t : \mathcal{E} \rightarrow \mathbb{R}_+$ ,  $\{i, j\} \mapsto t_{ij}$ . We can write the travelling salesman problem in the form of an integer linear programming problem, equivalent to  $(VC)$ ,

$$\begin{aligned} \min \sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij} \quad & \text{under the constraints} \\ x \in \{0, 1\}^{\mathcal{E}}, \\ \sum_{k \in \mathcal{V}, \{k,j\} \in \mathcal{E}} x_{kj} = 2, \quad & \text{for every } j \in \mathcal{V}, \\ x \text{ represents a connected graph with } |\mathcal{V}| \text{ edges covering all the vertices.} \end{aligned}$$

Let us distinguish a particular vertex  $1 \in \mathcal{V}$ . If, for every  $j \in \mathcal{V} \setminus \{1\}$ , we relax the constraints  $\sum_{k \in \mathcal{V}, \{k,j\} \in \mathcal{E}} x_{kj} = 2$ , we obtain the dual function  $\mathcal{D} : \mathbb{R}^{\mathcal{V} \setminus \{1\}} \rightarrow \mathbb{R}$

$$\begin{aligned} \mathcal{D}(\lambda) = \min \sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij} + \sum_{j \in \mathcal{V} \setminus \{1\}} \lambda_j (\sum_{k \in \mathcal{V}, \{k,j\} \in \mathcal{E}} x_{kj} - 2) \\ \text{under the constraints} \\ x \in \{0, 1\}^{\mathcal{E}} \\ \sum_{k \in \mathcal{V}, \{k,1\} \in \mathcal{E}} x_{k1} = 2 \\ x \text{ represents a connected graph with } |\mathcal{V}| \text{ edges covering all the vertices.} \end{aligned}$$

Up to a term  $-\sum_{j \in \mathcal{V} \setminus \{1\}} 2\lambda_j$ , we recognize in  $\mathcal{D}(\lambda)$  the minimum cost of a 1-tree for the graph  $\mathcal{D} = (\mathcal{V}, \mathcal{E})$ , where each edge  $\{i, j\}$  is equipped with the cost  $t_{ij} + \lambda_i + \lambda_j$ , for  $i, j \in \mathcal{V} \setminus \{1\}$ , and where each edge  $\{i, 1\}$ , for  $i \in \mathcal{V} \setminus \{1\}$ , has cost  $t_{1i} + \lambda_i$ . As we have already remarked in Section 11.6.1, the Kruskal algorithm of Section 11.5 makes it possible to calculate a 1-tree of minimum cost in time  $O(|\mathcal{E}| \log |\mathcal{E}|)$ , which allows us to efficiently implement the supergradient algorithm (11.71) to calculate the following bound:  $\sup_{\lambda \in \mathbb{R}^{\mathcal{V} \setminus \{1\}}} \mathcal{D}(\lambda)$ . This remarkable bound for the travelling

salesman problem, is due to Held and Karp. The 1-tree bound seen in Section 11.6.1 coincides with the value  $\mathcal{D}(0)$  obtained by not paying for the transgression of the constraints  $\sum_{k \in \mathcal{V}, \{k,j\} \in \mathcal{E}} x_{kj} = 2$ , for every  $j \in \mathcal{V} \setminus \{1\}$ .

**Exercise 11.6.2** Calculate the dual function  $\mathcal{D}(\lambda)$  for the problem (11.6.2), and thus recover the value of  $\max_{\lambda \in \mathbb{R}_+} \mathcal{D}(\lambda)$ .

**Exercise 11.6.3** Propose a Lagrangian relaxation giving an upper bound for the knapsack problem (11.52), and apply it to example (11.54).

**Exercise 11.6.4** The 'oriented' version of the travelling salesman problem consists of considering an oriented graph  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ , equipped with a time function  $t : \mathcal{A} \rightarrow \mathbb{R}_+$ , and of looking for an oriented circuit passing once and only once through each node. Propose a Lagrangian relaxation, by making it so that for each vector  $\lambda$  of Lagrange multipliers, the calculation of the dual function  $\mathcal{D}(\lambda)$  amounts to solving an assignment problem. Explain why this relaxation is less interesting than the bound of Held and Karp, in the particular case of the problem of the nonoriented travelling salesman problem.

**Exercise 11.6.5** We shall prove formula (11.70), in the case where all the  $J_i$  are convex functions from  $\mathbb{R}^n$  into  $\mathbb{R}$ . (The exclusion of the value  $+\infty$  is only a convenience, which will allow us to apply directly the results of Chapter 10, proved in the case of convex functions with finite values). Let us denote by  $C$  the right-hand side of (11.70). (1) Show that  $C \subset \partial J(x)$ . (2) We assume that  $0 \in \partial J(x)$ . By considering the convex programming problem

$$\min t \quad \text{under the constraints} \quad y \in \mathbb{R}^n, t \in \mathbb{R}, J_i(y) \leq t, \forall i \in I,$$

show that  $0 \in C$ . (3) Conclude that generally,  $C = \partial J(x)$ .

**Remark 11.6.6** Even though this is not the object of this course, let us say some words on the question of **complexity**. An algorithm is called **polynomial** (respectively **linear**) if its time of execution (on a normal computer) is bounded by a polynomial (respectively affine) function of the size of the data. We say that a problem is polynomial if it can be solved by a polynomial algorithm. The class P of polynomial problems formalizes the idea of a problem that we can solve 'well' in practice. For example, exercise 11.4.3 shows that the problem consisting of finding a path of minimal cost from one node to another, in a graph, is polynomial. Another example of a polynomial problem is the linear programming problem (the simplex algorithm can take an exponential time in some degenerate situations, but the interior point algorithms outlined in Section 11.2.3 finish in polynomial time). A class of combinatorial problems which seemed more difficult has received much attention. These are NP problems. Intuitively, a decision problem is NP if we can verify that a solution is correct in polynomial time. For example, the Hamiltonian Circuit problem which consists of deciding if there exists a circuit visiting each vertex of a graph once and only once, is an NP problem, since if you are given an arbitrary circuit, you can verify if it is Hamiltonian in polynomial time (it is enough to count the number of visits to each vertex). Another example of an NP problem is the Sat (satisfiability) problem, which consists of deciding if

a system of Boolean equations has a solution. A problem is called **NP-difficult** if it is at least as difficult as all the NP problems, which means that if we knew how to solve this problem in polynomial time, we would know how to solve all NP problems in polynomial time. The **NP-complete** problems are decision problems which are at the same time NP-difficult and in NP: the existence of such problems is a theorem due to Cook and Levin. For example, Sat and Hamiltonian Circuit are NP-complete problems. We say also that an optimization problem is NP-difficult when its decision version is NP-complete. For example, to find a Hamiltonian circuit of minimum cost, that is, a circuit for a travelling salesman of minimum cost, is an NP-difficult problem; the knapsack problem of exercise 11.4.5, or the shortest path problem with constraints of exercise 11.4.6, are also NP-difficult. We might think that NP-difficult problems are really more difficult than polynomial problems: to show that this is true, that is to show that  $P \neq NP$ , is a famous open problem. The reader interested in these questions can consult M.R. Garey and D.S. Johnson, *Computers and Intractability*, Freeman, San Francisco 1979, and also C.H. Papadimitriou, *Computational Complexity*, Addison-Wesley, Reading 1995. •

**Remark 11.6.7** To conclude this chapter of initiation into OR, let us mention the existence of two important approaches which are outside of the framework of this course. For problems which do not have good structure properties, or which are too large to apply exact methods, we often revert to **neighbourhood methods**: stochastic algorithms (whose most classical is **simulated annealing**, see, for example, [2]), or even **tabu search** [22]. A neighbourhood method is a heuristic which consists of exploring the space of solutions by successive modifications of an admissible solution. If we only accept the modifications which improve the criterion, we simply obtain a greedy heuristic, which can converge to a local minimum. Algorithms such as simulated annealing or tabu search specify how to manage modifications which temporarily degrade the criterion, and therefore arrive more often at a global optimum. A second very different general technique, (which often allows us to solve medium size problems optimally) is constraint programming, or CP, which intelligently explores the space of solutions using software which reduces the combinations by logical deductions. •

*This page intentionally left blank*

# Appendix Review of Hilbert spaces

---

We briefly give some properties of Hilbert spaces (for more details, we refer to [4], [25]). To simplify the presentation, we only consider the case of Hilbert spaces over  $\mathbb{R}$ .

**Definition 12.1.8** *A real Hilbert space is a vector space over  $\mathbb{R}$ , equipped with a scalar product, denoted  $\langle x, y \rangle$ , which is complete for the norm associated with this scalar product, denoted  $\|x\| = \sqrt{\langle x, x \rangle}$ . (We recall that a normed vector space is complete if every Cauchy sequence is a convergent sequence whose limit belongs to this space.)*

In everything that follows we denote by  $V$  a real Hilbert space, and  $\langle x, y \rangle$  its associated scalar product.

**Definition 12.1.9** *A set  $K \subset V$  is called convex if, for all  $x, y \in K$  and every real  $\theta \in [0, 1]$ , the element  $(\theta x + (1 - \theta)y)$  belongs to  $K$ .*

An essential result is the projection theorem over a convex set.

**Theorem 12.1.10 (projection over a convex set)** *Let  $V$  be a Hilbert space. Let  $K \subset V$  be a convex closed nonempty subset. For all  $x \in V$ , there exists a unique  $x_K \in K$  such that*

$$\|x - x_K\| = \min_{y \in K} \|x - y\|.$$

*Equivalently,  $x_K$  is characterized by the property*

$$x_K \in K, \quad \langle x_K - x, x_K - y \rangle \leq 0 \quad \forall y \in K. \quad (12.1)$$

*We call  $x_K$  the orthogonal projection over  $K$  of  $x$ .*



**Remark 12.1.11** Theorem 12.1.10 allows us to define a mapping  $P_K$ , called the projection operator on the convex set  $K$ , by setting  $P_K x = x_K$ . We easily verify that  $P_K$  is continuous and weakly contracting, that is to say that

$$\|P_K x - P_K y\| \leq \|x - y\| \quad \forall x, y \in V. \quad (12.2)$$

•

**Remark 12.1.12** A particular case of a convex closed  $K$  is a closed vector subspace  $W$ . In this case, the characterization (12.1) of  $x_W$  becomes

$$x_W \in W, \quad \langle x_W - x, z \rangle = 0 \quad \forall z \in W.$$

In effect, in (12.1) it is enough to take  $y = x_K \pm z$  with  $z$  arbitrary in  $W$ . •

**Proof.** Let  $y^n$  be a minimizing sequence, that is to say that  $y^n \in K$  satisfies

$$d_n = \|x - y^n\| \rightarrow d = \inf_{y \in K} \|x - y\| \quad \text{when } n \rightarrow +\infty.$$

Let us show that  $y^n$  is a Cauchy sequence. By using the symmetry of the scalar product, it becomes

$$\left\|x - \frac{1}{2}(y^n + y^p)\right\|^2 + \left\|\frac{1}{2}(y^n - y^p)\right\|^2 = \frac{1}{2}(d_n^2 + d_p^2).$$

Now, by the convexity of  $K$ ,  $(y^n + y^p)/2 \in K$ , and  $\|x - \frac{1}{2}(y^n + y^p)\|^2 \geq d^2$ . Consequently,

$$\|y^n - y^p\|^2 \leq 2(d_n^2 + d_p^2) - 4d^2,$$

which shows that  $y^n$  is a Cauchy sequence. As  $V$  is a Hilbert space, it is complete, therefore the sequence  $y^n$  converges to a limit  $x_K$ . In addition, as  $K$  is closed, this limit  $x_K$  belongs to  $K$ . Consequently, we have  $d = \|x - x_K\|$ . As every minimizing sequence is convergent, the limit is inevitably unique, and  $x_K$  is the only minimum point of  $\min_{y \in K} \|x - y\|$ .

Let  $x_K \in K$  be this minimum point. For all  $y \in K$  and  $\theta \in [0, 1]$ , by the convexity of  $K$ ,  $x_K + \theta(y - x_K)$  belongs to  $K$  and we have

$$\|x - x_K\|^2 \leq \|x - (x_K + \theta(y - x_K))\|^2.$$

By expanding the right-hand side, it becomes

$$\|x - x_K\|^2 \leq \|x - x_K\|^2 + \theta^2 \|y - x_K\|^2 - 2\theta \langle x - x_K, y - x_K \rangle,$$

which gives for  $\theta > 0$

$$0 \geq -2\langle x - x_K, y - x_K \rangle + \theta \|y - x_K\|^2.$$

By making  $\theta$  tend to 0, we obtain the characterization (12.1). Conversely, take  $x_K$  which satisfies this characterization. For all  $y \in K$  we have

$$\|x - y\|^2 = \|x - x_K\|^2 + \|x_K - y\|^2 + 2\langle x - x_K, x_K - y \rangle \geq \|x - x_K\|^2,$$

which proves that  $x_K$  is the orthogonal projection of  $x$  over  $K$ . □

**Definition 12.1.13** Let  $V$  be a Hilbert space for the scalar product  $\langle \cdot, \cdot \rangle$ . We say (countable) Hilbertian basis of  $V$  for a countable family  $(e_n)_{n \geq 1}$  of elements of  $V$  which is orthonormal for the scalar product and such that the vector space generated by this family is dense in  $V$ .

**Proposition 12.1.14** *Let  $V$  be a Hilbert space for the scalar product  $\langle \cdot, \cdot \rangle$ . Let  $(e_n)_{n \geq 1}$  be a Hilbertian basis of  $V$ . For every element  $x$  of  $V$ , there exists a unique sequence  $(x_n)_{n \geq 1}$  of real numbers such that the partial sum  $\sum_{n=1}^p x_n e_n$  converges to  $x$  when  $p$  tends to infinity, and this sequence is defined by  $x_n = \langle x, e_n \rangle$ . Further, we have*

$$\|x\|^2 = \langle x, x \rangle = \sum_{n \geq 1} |\langle x, e_n \rangle|^2. \quad (12.3)$$

We then write

$$x = \sum_{n \geq 1} \langle x, e_n \rangle e_n.$$

**Proof.** If there exists a sequence  $(x_n)_{n \geq 1}$  of real numbers such that  $\lim_{p \rightarrow +\infty} \sum_{n=1}^p x_n e_n = x$ , then by projection over  $e_n$  (and as this sequence is by definition independent of  $p$ ) we have  $x_n = \langle x, e_n \rangle$ , which proves the uniqueness of the sequence  $(x_n)_{n \geq 1}$ . Let us now show its existence. By the definition of a Hilbertian basis, for all  $x \in V$  and for all  $\epsilon > 0$ , there exists  $y$ , a finite linear combination of  $(e_n)_{n \geq 1}$ , such that  $\|x - y\| < \epsilon$ . Thanks to theorem 12.1.10 we can define a linear mapping  $S_p$  which, for every point  $z \in V$ , produces  $S_p z = z_W$ , where  $z_W$  is the orthogonal projection over the vector subspace  $W$  generated by the first  $p$  vectors  $(e_n)_{1 \leq n \leq p}$ . From (12.1),  $(z - S_p z)$  is orthogonal to every element of  $W$ , therefore in particular to  $S_p z$ . We deduce that

$$\|z\|^2 = \|z - S_p z\|^2 + \|S_p z\|^2, \quad (12.4)$$

which implies

$$\|S_p z\| \leq \|z\| \quad \forall z \in V.$$

As  $S_p z$  is generated by the  $(e_n)_{1 \leq n \leq p}$ , and  $(z - S_p z)$  is orthogonal to each of the  $(e_n)_{1 \leq n \leq p}$ , we easily see that

$$S_p z = \sum_{n=1}^p \langle z, e_n \rangle e_n.$$

For sufficiently large  $p$ , we have  $S_p y = y$  as  $y$  is a finite linear combination of  $(e_n)_{n \geq 1}$ . Consequently,

$$\|S_p x - x\| \leq \|S_p(x - y)\| + \|y - x\| \leq 2\|x - y\| \leq 2\epsilon.$$

We deduce the convergence of  $S_p x$  to  $x$ . From this convergence and from equation (12.4) we have

$$\lim_{p \rightarrow +\infty} \|S_p x\|^2 = \|x\|^2,$$

which is none other than the Parseval summation formula (12.3).  $\square$

The existence of a countable Hilbertian basis is not guaranteed for all Hilbert spaces. The following proposition gives a necessary and sufficient condition for the existence of a countable Hilbertian basis.

**Proposition 12.1.15** *Let  $V$  be a separable Hilbert space (that is, there exists a countable family which is dense in  $V$ ). Then there exists a countable Hilbertian basis of  $V$ .*

**Proof.** Let  $(v_n)_{n \geq 1}$  be the family whose generated vector subspace is dense in  $V$  (even if it means renumbering the  $v_n$  and removing some of them, we can always assume that they are linearly independent). By an application of the Gram–Schmidt procedure to this family, we obtain an orthonormal family  $(e_n)_{n \geq 1}$ . As  $[v_1, \dots, v_n] = [e_1, \dots, e_n]$ , we deduce that the vector subspace generated by  $(e_n)_{n \geq 1}$  coincides with that generated by  $(v_n)_{n \geq 1}$  which is dense in  $V$ . Therefore,  $(e_n)_{n \geq 1}$  is a Hilbertian basis.  $\square$

**Definition 12.1.16** Let  $V$  and  $W$  be two real Hilbert spaces. A linear mapping  $A$  from  $V$  into  $W$  is called *continuous* if there exists a constant  $C$  such that

$$\|Ax\|_W \leq C\|x\|_V \quad \forall x \in V.$$

The smallest constant  $C$  which satisfies this inequality is the norm of the linear mapping  $A$ , in other words

$$\|A\| = \sup_{x \in V, x \neq 0} \frac{\|Ax\|_W}{\|x\|_V}.$$

Often, we will use the equivalent notation of the operator instead of mapping between Hilbert spaces (we will speak of a continuous linear operator rather than a continuous linear mapping). If  $V$  is finite dimensional, then all the linear mappings from  $V$  into  $W$  are continuous, but this is no longer true if  $V$  is infinite dimensional.

**Definition 12.1.17** Let  $V$  be a real Hilbert space. Its dual  $V'$  is the set of **continuous** linear forms over  $V$ , that is to say, the set of continuous linear mappings from  $V$  into  $\mathbb{R}$ . By definition, the norm of an element  $L \in V'$  is

$$\|L\|_{V'} = \sup_{x \in V, x \neq 0} \frac{|L(x)|}{\|x\|}.$$

In a Hilbert space, duality has a very simple interpretation thanks to the Riesz representation theorem which allows us to identify a Hilbert space with its dual by isomorphism.

**Theorem 12.1.18 (Riesz representation)** Let  $V$  be a real Hilbert space, and let  $V'$  be its dual. For every continuous linear form  $L \in V'$  there exists a unique  $y \in V$  such that

$$L(x) = \langle y, x \rangle \quad \forall x \in V.$$

Further, we have  $\|L\|_{V'} = \|y\|$ .

**Proof.** Let  $M = \text{Ker } L$ . This is a closed subspace of  $V$  since  $L$  is continuous. If  $M = V$ , then  $L$  is identically zero and we have  $y = 0$ . If  $M \neq V$ , then there exists  $z \in V \setminus M$ . Let  $z_M \in M$  be its projection over  $M$ . As  $z$  does not belong to  $M$ ,  $z - z_M$  is nonzero and, by theorem 12.1.10, is orthogonal to every element of  $M$ . Finally, let

$$z_0 = \frac{z - z_M}{\|z - z_M\|}.$$

Every vector  $x \in V$  can be written

$$x = w + \lambda z_0 \quad \text{with } \lambda = \frac{L(x)}{L(z_0)}.$$

We see easily that  $L(w) = 0$ , therefore  $w \in M$ . This proves that  $V = \text{Vect}(z_0) \oplus M$ . By definition of  $z_M$  and of  $z_0$ , we have  $\langle w, z_0 \rangle = 0$ , which implies

$$L(x) = \langle x, z_0 \rangle L(z_0),$$

from where then we have the result with  $y = L(z_0)z_0$  (the uniqueness is obvious). On the other hand, we have

$$\|y\| = |L(z_0)|,$$

and

$$\|L\|_{V'} = \sup_{x \in V, x \neq 0} \frac{|L(x)|}{\|x\|} = L(z_0) \sup_{x \in V, x \neq 0} \frac{\langle x, z_0 \rangle}{\|x\|}.$$

The maximum in the last term of this equality is attained by  $x = z_0$ , which implies that  $\|L\|_{V'} = \|y\|$ .  $\square$

An essential result to prove the Farkas lemma 10.2.17 (used in optimization) is the following geometrical property which is completely in agreement with intuition.

**Theorem 12.1.19 (separation of a point and a convex set)** *Let  $K$  be a closed convex nonempty set of a Hilbert space  $V$ , and  $x_0 \notin K$ . Then there exists a closed hyperplane of  $V$  which separates  $x_0$  and  $K$  strictly, that is to say, there exists a linear form  $L \in V'$  and  $\alpha \in \mathbb{R}$  such that*

$$L(x_0) < \alpha < L(x) \quad \forall x \in K. \quad (12.5)$$

**Proof.** Let us denote by  $x_K$  the projection of  $x_0$  over  $K$ . Since  $x_0 \notin K$ , we have  $x_K - x_0 \neq 0$ . Let  $L$  be the linear form defined for all  $y \in V$  by  $L(y) = \langle x_K - x_0, y \rangle$ , and let  $\alpha = (L(x_K) + L(x_0))/2$ . From (12.1), we have  $L(x) \geq L(x_K) > \alpha > L(x_0)$  for all  $x \in K$ , which finishes the proof.  $\square$

We finally need, to prove the Minkowski theorem 11.3.1, a variant of the separation theorem, involving the important notion of hyperplane of support. If  $K$  is a convex set of a Hilbert space  $V$ , we say **hyperplane of support** of  $K$  at a point  $x$  to mean an affine hyperplane  $H = \{y \in V \mid L(y) = \alpha\}$ , with  $L \in V'$ ,  $L \neq 0$ , and  $\alpha \in \mathbb{R}$ , such that  $\alpha = L(x) \leq L(y)$ , for all  $y \in C$ .

**Corollary 12.1.20 (hyperplane of support)** *There exists a hyperplane of support at every boundary point of a convex closed set  $K$  of a finite dimensional Hilbert space.*

**Proof.** Let  $x$  be a boundary point of  $K$ : there then exists a sequence  $x_n \in V \setminus K$ , with  $x_n \rightarrow x$ . The separation theorem 12.1.19 gives, for all  $n$ , a nonzero linear form  $L_n$  such that  $L_n(x_n) \leq L_n(y)$  for all  $y \in K$ . We can choose  $L_n$  with norm 1. As  $V$  is finite dimensional, the unit sphere of  $V'$  is compact, and replacing  $L_n$  by a subsequence, we can assume that  $L_n$  converges to a linear form  $L$ , which is nonzero, since it has norm 1. It is enough now to pass to the limit in  $L_n(x_n) \leq L_n(y)$ , which we justify by writing  $L_n(x_n) = L_n(x_n - x) + L_n(x)$  and by noting that  $|L_n(x_n - x)| \leq \|L_n\| \|x_n - x\| = \|x_n - x\|$ , to obtain  $L(x) \leq L(y)$  no matter what  $y \in K$ . Thus,  $H = \{y \in V \mid L(y) = L(x)\}$  is a hyperplane of support of  $K$  at  $x$ .  $\square$

**Remark 12.1.21** The proof of corollary 12.1.20 does not extend to infinite dimensions: in this case, the sequence  $L_n$  has a value of adherence  $L$  for the weak topology, but nothing says that  $L \neq 0$ . As a counterexample, let us consider the set  $K$  of sequences of  $\ell_2$  with positive or zero terms, which is a convex closed set of  $\ell_2$  with empty interior. Every point of  $K$  is therefore a boundary point, but if  $x$  is a sequence of  $\ell_2$  with strictly positive terms, there does not exist a hyperplane of support at  $x$ .  $\bullet$

# Appendix Matrix Numerical Analysis

---

This appendix is dedicated to the numerical analysis of matrix calculations and more precisely to algorithms used to solve linear systems (particularly those coming from the finite element method), and to calculating the eigenvalues and eigenvectors of a self-adjoint matrix (occurring in the calculation of the eigenmodes of a mechanical model). For more details we refer to the works [8] and [24].

## 13.1 Solution of linear systems

We say linear system for the problem which consists of finding the solutions  $x \in \mathbb{R}^n$  (if they exist) of the following algebraic equation

$$Ax = b, \tag{13.1}$$

where  $A$  belongs to the set  $\mathcal{M}_n(\mathbb{R})$  of real square matrices of order  $n$ , and  $b \in \mathbb{R}^n$  is a right-hand side vector. Of course, we have the well-known Cramer formulas at our disposal which, for an invertible matrix  $A$  with columns  $(a^1, \dots, a^n)$ , gives the solution of (13.1) in its components

$$x_i = \frac{\det(a^1, \dots, a^{i-1}, b, a^{i+1}, \dots, a^n)}{\det A}.$$

We might believe that this explicit formula is enough for our needs. But it is not, as the Cramer formulas are **completely unsuited** to calculating efficiently the solution of a linear system. Indeed, their cost in execution time on a computer is prohibitive: we must calculate  $n+1$  determinants and if we use the method of expansion by row (or column) each determinant needs more than  $n!$  multiplications. In total, the Cramer method therefore needs more than  $(n+1)!$  multiplications, which is unimaginable: for example, for  $n = 50$ , if the calculations are carried out on a computer operating at 1 Giga flop (a billion operations per second), the time of calculation is of the order of  $4.8 \times 10^{49}$  years! Even if we use a better method to calculate the determinants, the Cramer method is not competitive compared with the algorithms that we shall see (where the number of operations will typically be of the order of  $n^3$ ).

We shall see two types of methods for the solution of linear systems: those called direct, that is to say which allow us to calculate the exact solution in a finite number of operations, and those called **iterative**, that is to say which calculate a sequence of approximate solutions which converge to the exact solution.

### 13.1.1 Review of matrix norms

We start by recalling the idea of **subordinate norm** for matrices. We denote by  $\mathcal{M}_n(\mathbb{R})$  (respectively  $\mathcal{M}_n(\mathbb{C})$ ) the set of real (respectively complex) square matrices of order  $n$ . Even if we consider real matrices, it is necessary, for technical reasons which will be seen in remark 13.1.4, to consider complex matrices.

**Definition 13.1.1** Let  $\|\cdot\|$  be a vector norm on  $\mathbb{C}^n$ . We associate with it a matrix norm, called subordinate to this vector norm, defined by

$$\|A\| = \sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

By abuse of language we use the same notation for vector norms and subordinate matrix norms. We easily verify that a subordinate norm defined in this way is a matrix norm over  $\mathcal{M}_n(\mathbb{C})$  or  $\mathcal{M}_n(\mathbb{R})$ .

**Lemma 13.1.2** Let  $\|\cdot\|$  be a subordinate matrix norm over  $\mathcal{M}_n(\mathbb{C})$ .

1. For every matrix  $A$ , the norm  $\|A\|$  is also defined by

$$\|A\| = \sup_{x \in \mathbb{C}^n, \|x\|=1} \|Ax\| = \sup_{x \in \mathbb{C}^n, \|x\| \leq 1} \|Ax\|.$$

2. There exists  $x_A \in \mathbb{C}^n, x_A \neq 0$  such that  $\|A\| = \frac{\|Ax_A\|}{\|x_A\|}$ .

3. The identity matrix satisfies  $\|I\| = 1$ .

4. Let  $A$  and  $B$  be two matrices. We have  $\|AB\| \leq \|A\| \|B\|$ .

**Proof.** The first point is obvious. The second can be shown by remarking that the continuous function  $\|Ax\|$  attains its maximum on the compact set  $\{x \in \mathbb{C}^n, \|x\| = 1\}$ . The third is obvious, while the fourth is a consequence of the inequality  $\|ABx\| \leq \|A\| \|Bx\|$ .  $\square$

**Remark 13.1.3** There exist matrix norms which are not subordinate to any vector norm. The best example is the Euclidean norm defined by  $\|A\| = \sqrt{\sum_{i,j=1}^n |a_{ij}|^2}$ . Indeed, we have  $\|I\| = \sqrt{n}$ , which is not possible for a subordinate norm.  $\bullet$

We denote by  $\|A\|_p$  the matrix norm subordinate to the vector norm on  $\mathbb{C}^n$  defined for  $p \geq 1$  by  $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ , and for  $p = +\infty$  by  $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$ . We can calculate explicitly some of these subordinate norms. (In everything that follows we denote by  $A^*$  the adjoint matrix of  $A$ .)

**Exercise 13.1.1** Show that

1.  $\|A\|_2 = \|A^*\|_2 = \text{maximum singular values of } A$ ,
2.  $\|A\|_1 = \max_{1 \leq j \leq n} (\sum_{i=1}^n |a_{ij}|)$ ,
3.  $\|A\|_\infty = \max_{1 \leq i \leq n} (\sum_{j=1}^n |a_{ij}|)$ .

**Remark 13.1.4** A real matrix can be considered to be a matrix of  $\mathcal{M}_n(\mathbb{R})$ , or a matrix of  $\mathcal{M}_n(\mathbb{C})$  as  $\mathbb{R} \subset \mathbb{C}$ . If  $\|\cdot\|_{\mathbb{C}}$  is a vector norm in  $\mathbb{C}^n$ , we can define its restriction  $\|\cdot\|_{\mathbb{R}}$  to  $\mathbb{R}^n$  which is also a vector norm in  $\mathbb{R}^n$ . For a real matrix  $A \in \mathcal{M}_n(\mathbb{R})$ , we can therefore define two subordinate matrix norms  $\|A\|_{\mathbb{C}}$  and  $\|A\|_{\mathbb{R}}$  by

$$\|A\|_{\mathbb{C}} = \sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\|Ax\|_{\mathbb{C}}}{\|x\|_{\mathbb{C}}} \quad \text{and} \quad \|A\|_{\mathbb{R}} = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_{\mathbb{R}}}{\|x\|_{\mathbb{R}}}.$$

*A priori* these two definitions can be distinct. Thanks to the explicit formulas of exercise 13.1.1, we know that they coincide if  $\|x\|_{\mathbb{C}}$  is one of the norms  $\|x\|_1$ ,  $\|x\|_2$ , or  $\|x\|_\infty$ . However, for other vector norms we can have  $\|A\|_{\mathbb{C}} > \|A\|_{\mathbb{R}}$ . In addition, in the proof of proposition 13.1.7 we need the definition over  $\mathbb{C}$  of the subordinate norm even if the matrix is real. This is why we use  $\mathbb{C}$  in the definition 13.1.1 of the subordinate norm. •

**Definition 13.1.5** Let  $A$  be a matrix in  $\mathcal{M}_n(\mathbb{C})$ . We say spectral radius of  $A$ , denoted by  $\rho(A)$ , for the maximum of the modulus of the eigenvalues of  $A$ .

The spectral radius  $\rho(A)$  is not a norm over  $\mathcal{M}_n(\mathbb{C})$ . Indeed, we can have  $\rho(A) = 0$  with  $A \neq 0$  (take, for example, a triangular matrix with zeros on the diagonal). However, the lemma below shows that this is a norm on the set of normal matrices.

**Lemma 13.1.6** If  $U$  is a unitary matrix ( $U^* = U^{-1}$ ), we have  $\|UA\|_2 = \|AU\|_2 = \|A\|_2$ . Consequently, if  $A$  is a normal matrix ( $A^*A = AA^*$ ), then  $\|A\|_2 = \rho(A)$ .

**Proof.** As  $U^*U = I$ , we have

$$\|UA\|_2^2 = \sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\|UAx\|_2^2}{\|x\|_2^2} = \sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\langle U^*UAx, Ax \rangle}{\langle x, x \rangle} = \|A\|_2^2.$$

On the other hand, the change of variable  $y = Ux$  satisfies  $\|x\|_2 = \|y\|_2$ , and therefore

$$\|AU\|_2^2 = \sup_{x \in \mathbb{C}^n, x \neq 0} \frac{\|AUx\|_2^2}{\|x\|_2^2} = \sup_{y \in \mathbb{C}^n, y \neq 0} \frac{\|Ay\|_2^2}{\|U^{-1}y\|_2^2} = \sup_{y \in \mathbb{C}^n, y \neq 0} \frac{\|Ay\|_2^2}{\|y\|_2^2} = \|A\|_2^2.$$

If  $A$  is normal, it is diagonalizable in an orthonormal basis of eigenvectors and we deduce from the preceding results that  $\|A\|_2 = \|\text{diag}(\lambda_i)\|_2 = \rho(A)$ . □

We now compare the norm of a matrix  $A$  with its spectral radius  $\rho(A)$ .



**Proposition 13.1.7** *Let  $\|\cdot\|$  be a subordinate norm over  $\mathcal{M}_n(\mathbb{C})$ . We have*

$$\rho(A) \leq \|A\|.$$

*Conversely, for every matrix  $A$  and for all real numbers  $\epsilon > 0$ , there exists a subordinate norm  $\|\cdot\|$  (which depends on  $A$  and  $\epsilon$ ) such that*

$$\|A\| \leq \rho(A) + \epsilon. \quad (13.2)$$

**Proof.** Let  $\lambda \in \mathbb{C}$  be an eigenvalue of  $A$  such that  $\rho(A) = |\lambda|$ , and  $x^0 \neq 0$  an associated eigenvector ( $Ax^0 = \lambda x^0$ ). We have

$$\|\lambda x^0\| = \rho(A)\|x^0\| = \|Ax^0\| \leq \|A\|\|x^0\|,$$

from which we deduce  $\rho(A) \leq \|A\|$ . As the eigenvector  $x^0$  can be complex, it is essential to use a vector norm on  $\mathbb{C}^n$  even for a real matrix (cf. remark 13.1.4). Conversely, there exists an invertible matrix  $U$  such that  $T = U^{-1}AU$  is upper triangular. For all  $\delta > 0$  we define a diagonal matrix  $D_\delta = \text{diag}(1, \delta, \delta^2, \dots, \delta^{n-1})$  such that the matrix  $T_\delta$  defined by

$$T_\delta = (UD_\delta)^{-1}A(UD_\delta) = D_\delta^{-1}TD_\delta$$

satisfies

$$T_\delta = \begin{pmatrix} t_{11} & \delta t_{12} & \cdots & \delta^{n-1}t_{1n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \delta t_{n-1n} \\ 0 & \cdots & 0 & t_{nn} \end{pmatrix} \quad \text{with} \quad T = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & t_{nn} \end{pmatrix}.$$

Given  $\epsilon > 0$ , we can choose  $\delta$  sufficiently small so that the off-diagonal elements of  $T_\delta$  are also very small, for example, so that for all  $1 \leq i \leq n-1$ ,

$$\sum_{j=i+1}^n \delta^{j-i}|t_{ij}| \leq \epsilon.$$

Then the mapping  $B \rightarrow \|(UD_\delta)^{-1}B(UD_\delta)\|_\infty$  is a subordinate norm (which depends on  $A$  and  $\epsilon$ ) which satisfies (13.2).  $\square$

**Lemma 13.1.8** *Let  $A$  be a matrix of  $\mathcal{M}_n(\mathbb{C})$ . The following four conditions are equivalent*

1.  $\lim_{i \rightarrow +\infty} A^i = 0$ ,
2.  $\lim_{i \rightarrow +\infty} A^i x = 0$  for every vector  $x \in \mathbb{C}^n$ ,
3.  $\rho(A) < 1$ ,
4. there exists at least one subordinate matrix norm such that  $\|A\| < 1$ .

**Proof.** Let us show first of all that (1) implies (2). The inequality

$$\|A^i x\| \leq \|A^i\| \|x\|$$

shows that  $\lim_{i \rightarrow +\infty} A^i x = 0$ . Then, (2) implies (3) as, if  $\rho(A) \geq 1$ , then there exists  $\lambda$  and  $x \neq 0$  such that  $Ax = \lambda x$  and  $|\lambda| = \rho(A)$ , and, consequently, the sequence  $A^i x = \lambda^i x$  cannot converge to 0. As ‘(3) implies (4)’ is an immediate consequence of proposition 13.1.7, it only remains to show that (4) implies (1). For this, we consider the subordinate matrix norm such that  $\|A\| < 1$ , and we have

$$\|A^i\| \leq \|A\|^i \rightarrow 0 \quad \text{when } i \rightarrow +\infty,$$

which shows that  $A^i$  tends to 0. □

### 13.1.2 Conditioning and stability

Before describing the algorithms for the solution of linear systems, we must consider the problems of precision and stability due to rounding errors. Indeed, in a computer there are no exact calculations, and the precision is limited because of the number of bits used to represent real numbers: usually 32 or 64 bits (which makes about 8 or 16 significant figures). We must therefore pay great attention to the inevitable rounding errors and to their propagation throughout a calculation. Numerical methods for the solution of linear systems which do not amplify these errors are called **stable**. In practice, we shall therefore use algorithms which are both **efficient and stable**. This amplification of errors depends on the matrix considered. To quantify this phenomenon, we introduce the idea of the condition number of a matrix.

**Definition 13.1.9** *Take a subordinate matrix norm that we denote by  $\|A\|$  (see definition 13.1.1). We say the condition number of a matrix  $A \in \mathcal{M}_n(\mathbb{C})$ , relative to this norm, for the value defined by*

$$\text{cond}(A) = \|A\| \cdot \|A^{-1}\|.$$

This idea of condition number will allow us to measure the amplification of the errors in the data (right-hand side or matrix) which result.

**Proposition 13.1.10** *Let  $A$  be an invertible matrix. Let  $b \neq 0$  be a nonzero vector.*

1. *Let  $x$  and  $x + \delta x$  be the respective solutions of the systems*

$$Ax = b \quad \text{and} \quad A(x + \delta x) = b + \delta b.$$

*Then we have*

$$\frac{\|\delta x\|}{\|x\|} \leq \text{cond}(A) \frac{\|\delta b\|}{\|b\|}. \quad (13.3)$$

2. Let  $x$  and  $x + \delta x$  be the respective solutions of the systems

$$Ax = b \quad \text{and} \quad (A + \delta A)(x + \delta x) = b.$$

Then we have

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \text{cond}(A) \frac{\|\delta A\|}{\|A\|}. \quad (13.4)$$

Moreover, these inequalities are optimal.

**Remark 13.1.11** We shall say that a matrix is well conditioned if its condition number is close to 1 (its minimal value) and that it is ill conditioned if its condition number is large. Because of the results of proposition 13.1.10, in practice it will be necessary to pay attention to rounding errors if we solve a linear system for an ill-conditioned matrix. •

**Proof.** To show the first result, we remark that  $A\delta x = \delta b$ , and therefore  $\|\delta x\| \leq \|A^{-1}\| \cdot \|\delta b\|$ . However, we also have  $\|b\| \leq \|A\| \|x\|$ , which gives (13.3). This inequality is optimal in the following sense: for every matrix  $A$ , there exists  $\delta b$  and  $x$  (which depend on  $A$ ) such that (13.3) is in fact an equality. Indeed, from a property of subordinate matrix norms (see lemma 13.1.2) there exists  $x$  such that  $\|b\| = \|A\| \|x\|$  and there exists  $\delta b$  such that  $\|\delta x\| = \|A^{-1}\| \|\delta b\|$ .

To obtain (13.4) we remark that  $A\delta x + \delta A(x + \delta x) = 0$ , and therefore  $\|\delta x\| \leq \|A^{-1}\| \|\delta A\| \|x + \delta x\|$ , which implies (13.4). To prove the optimality, we shall show that for every matrix  $A$  there exists a perturbation  $\delta A$  and a right-hand side  $b$  for which there is equality. Thanks to lemma 13.1.2 there exists  $y \neq 0$  such that  $\|A^{-1}y\| = \|A^{-1}\| \|y\|$ . Let  $\epsilon$  be a nonzero scalar. We set  $\delta A = \epsilon I$  and  $b = (A + \delta A)y$ . We verify then that  $y = y + \delta x$  and  $\delta x = -\epsilon A^{-1}y$ , and as  $\|\delta A\| = |\epsilon|$  we obtain the equality in (13.4). □

The condition numbers most used in practice are

$$\text{cond}_p(A) = \|A\|_p \|A^{-1}\|_p \quad \text{for } p = 1, 2, +\infty,$$

where the norms  $\|A\|_p$  are explicitly defined in lemma 13.1.1. We easily verify a certain number of properties of condition number.

**Exercise 13.1.2** Take a matrix  $A \in \mathcal{M}_n(\mathbb{C})$ . Verify that

- (1)  $\text{cond}(A) = \text{cond}(A^{-1}) \geq 1, \quad \text{cond}(\alpha A) = \text{cond}(A) \quad \forall \alpha \neq 0,$
- (2) for an arbitrary matrix,  $\text{cond}_2(A) = \mu_n(A)/\mu_1(A)$ , where  $\mu_1(A), \mu_n(A)$  are respectively the smallest and the largest singular value of  $A$ ,
- (3) for a normal matrix,  $\text{cond}_2(A) = |\lambda_n(A)|/|\lambda_1(A)|$ , where  $|\lambda_1(A)|, |\lambda_n(A)|$  are respectively the smallest and the largest eigenvalue in modulus of  $A$ ,

- (4) for every unitary matrix  $U$ ,  $\text{cond}_2(U) = 1$ ,  
 (5) for every unitary matrix  $U$ ,  $\text{cond}_2(AU) = \text{cond}_2(UA) = \text{cond}_2(A)$ .

**Exercise 13.1.3** Show that the condition number of the stiffness matrix  $\mathcal{K}_h$ , given by (6.12) for the  $P_1$  finite element method applied to the Laplacian, is

$$\text{cond}_2(\mathcal{K}_h) \approx \frac{4}{\pi^2 h^2}. \quad (13.5)$$

We shall show that the eigenvalues of  $\mathcal{K}_h$  are

$$\lambda_k = 4h^{-2} \sin^2 \left( \frac{k\pi}{2(n+1)} \right) \quad 1 \leq k \leq n,$$

for eigenvectors  $u^k$

$$u_j^k = \sin \left( \frac{jk\pi}{n+1} \right) \quad 1 \leq j, \quad k \leq n.$$

**Remark 13.1.12** The estimate (13.5) of the condition number of the stiffness matrix  $\mathcal{K}_h$  seems very pessimistic, even catastrophic. Indeed, the finite element method converges if  $h = 1/(n+1)$  tends to zero. In other words, precise results can only be obtained if the matrix is very large and very ill conditioned. But in this case, the inevitable rounding errors on the right-hand side or on the matrix risk being amplified enormously to the point of making the discrete solution  $u_h$  very different from its predicted limit. Very fortunately, this does not happen in practice as the right-hand side  $b_h$  of the linear system  $\mathcal{K}_h U_h = b_h$  is not arbitrary and does not make the inequalities of proposition 13.1.10 optimal. If we return to the proof of the optimality of these inequalities, we realize that it is obtained for a vector  $b$  which is an eigenvector of  $\mathcal{K}_h$  associated with its largest eigenvalue  $\lambda_n$ . From exercise 13.1.3, such an eigenvector oscillates strongly on the mesh (its components change sign from one element to the next). If the right-hand side  $b_h$  is more ‘regular’ (that is to say that it is a linear combination of the  $K$  first eigenvectors  $u^k$  of  $\mathcal{K}_h$ ), we can improve the result of proposition 13.1.10 by obtaining

$$\frac{\|\delta x\|}{\|x\|} \leq C(K) \frac{\|\delta b\|}{\|b\|},$$

where  $C(K)$  is a constant independent of  $n$ . This is exactly what happens in practice, and this last inequality justifies the use of the finite element method despite the presence of rounding errors in calculations on computers. •

### 13.1.3 Direct methods

#### Gaussian elimination

The principal idea of this method is to reduce the problem to the solution of a linear system whose matrix is triangular. Indeed, the solution of a linear system,  $Tx = b$ ,

where the matrix  $T$  is triangular and invertible, is very easy by simple recursive substitution. Indeed, the system

$$\left\{ \begin{array}{cccccc} t_{1,1}x_1 + & t_{1,2}x_2 + & \cdots & & \cdots & t_{1,n}x_n = b_1 \\ & t_{2,2}x_2 + & \ddots & & \cdots & t_{2,n}x_n = b_2 \\ & & \ddots & & \ddots & \vdots \\ & & & t_{n-1,n-1}x_{n-1} + & t_{n-1,n}x_n = b_{n-1} \\ & & & & t_{n,n}x_n = b_n \end{array} \right.$$

is solved by first calculating  $x_n = b_n/t_{n,n}$ , then  $x_{n-1}$ , and so on until  $x_1$ . We call this procedure **back substitution** (in the case of a lower triangular matrix, the similar procedure which calculates the components of the solution from  $x_1$  to  $x_n$  is called **forward substitution**). Let us remark that we have therefore solved the system  $Tx = b$  without inverting the matrix  $T$ . In the same way, the method of Gaussian elimination will solve the system  $Ax = b$  without calculating the inverse of the matrix  $A$ .

Gaussian elimination decomposes into three steps:

- (1) elimination: calculation of an invertible matrix  $M$  such that  $MA = T$  is upper triangular,
- (2) update of the right-hand side: simultaneous calculation of  $Mb$ ,
- (3) substitution: solution of the triangular system  $Tx = Mb$  by simple back substitution.

The existence of such a matrix  $M$  is guaranteed by the following result for which we will give a constructive proof which is none other than Gaussian elimination.

**Proposition 13.1.13** *Let  $A$  be a square matrix (invertible or not). There exists at least an invertible matrix  $M$  such that the matrix  $T = MA$  is upper triangular.*

**Proof.** The principle is to construct a sequence of matrices  $A^k$ ,  $1 \leq k \leq n$ , whose first  $(k-1)$  columns are filled with zeros under the diagonal. By successive modifications, we go from  $A^1 = A$  to  $A^n = T$  which is upper triangular. We denote by  $(a_{ij}^k)_{1 \leq i, j \leq n}$  the elements of the matrix  $A^k$ , and say pivot of  $A^k$  for the element  $a_{kk}^k$ . To go from the matrix  $A^k$  to the matrix  $A^{k+1}$ , we must be sure first of all that the pivot  $a_{kk}^k$  is not zero. If it is, we permute the  $k$ th row with another row to put a nonzero element in the position of the pivot. Then we proceed to the elimination of all the elements of the  $k$ th column below the  $k$ th row by making linear combinations of the current row with the  $k$ th row.

More precisely, we carry out the following operations. We multiply  $A^k$  by a permutation matrix  $P^k$  to obtain  $\tilde{A}^k = P^k A^k$  such that its pivot  $\tilde{a}_{kk}^k$  is nonzero. If  $a_{kk}^k \neq 0$ , then it is enough to take  $P^k = I$ . Otherwise, if there exists  $a_{ik}^k \neq 0$  with  $i \geq k+1$ , we permute the  $k$ th row with the  $i$ th by taking  $P^k = (e_1, \dots, e_{k-1}, e_i, e_{k+1}, \dots, e_{i-1}, e_k, e_{i+1}, \dots, e_n)$  (if all

the elements of the  $k$ th column under the diagonal,  $a_{ik}^k$  with  $i \geq k$ , are zero, then there is nothing to do!). Then we multiply  $\tilde{A}^k$  by a matrix  $E^k$ , defined by

$$E^k = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -\frac{\tilde{a}_{k+1,k}^k}{\tilde{a}_{k,k}^k} & 1 & & \\ & & \vdots & & \ddots & \\ 0 & & -\frac{\tilde{a}_{n,k}^k}{\tilde{a}_{k,k}^k} & & & 1 \end{pmatrix}, \quad (13.6)$$

which eliminates all the coefficients of the  $k$ th column below the diagonal. We set

$$A^{k+1} = E^k \tilde{A}^k = \begin{pmatrix} \tilde{a}_{11}^1 & \cdots & \cdots & \cdots & \cdots & \tilde{a}_{1n}^1 \\ 0 & \ddots & & & & \vdots \\ \vdots & \ddots & \tilde{a}_{k,k}^k & \tilde{a}_{k,k+1}^k & \cdots & \tilde{a}_{k,n}^k \\ \vdots & & 0 & a_{k+1,k+1}^{k+1} & \cdots & a_{k+1,n}^{k+1} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n,k+1}^{k+1} & \cdots & a_{n,n}^{k+1} \end{pmatrix}$$

with  $a_{ij}^{k+1} = \tilde{a}_{ij}^k - (\tilde{a}_{i,k}^k / \tilde{a}_{k,k}^k) \tilde{a}_{k,j}^k$  for  $k+1 \leq i, j \leq n$ . The matrix  $A^{k+1}$  therefore has the desired form with its  $k$  first columns having zeros under the diagonal. After  $(n-1)$  steps, the matrix  $A^n$  is upper triangular and satisfies  $A^n = MA$  with  $M = E^{n-1}P^{n-1} \cdots E^1P^1$ . The matrix  $M$  is invertible as  $\det P^i = \pm 1$  and  $\det E^i = 1$ .

We can update the right-hand side (that is to say calculate  $Mb$ ) as we calculate the matrices  $P^k$  and  $E^k$ . We construct a sequence of right-hand sides  $(b^k)_{1 \leq k \leq n}$  defined by

$$b^1 = b, \quad b^{k+1} = E^k P^k b^k \quad \text{for } 1 \leq k \leq n-1,$$

and we have  $b^n = Mb$ . To solve the linear system  $Ax = b$  it only remains to solve the system  $A^n x = Mb$  where  $A^n = T$  is an upper triangular matrix.  $\square$

**Remark 13.1.14** Let us mention some practical aspects of the Gaussian elimination method.

1. We never calculate  $M$ ! We do not need to multiply the matrices  $E^k$  and  $P^k$  to calculate  $Mb$  and  $A^n$ .
2. If  $A$  is not invertible, one of the diagonal coefficients of  $A^n = T$  is zero and we will not be able to solve  $Tx = Mb$ . Conversely, the elimination is always possible.
3. At step  $k$ , we only modify the part of the rows  $k+1$  to  $n$  between the columns  $k+1$  to  $n$ .
4. As a by-product of Gaussian elimination, we can calculate the determinant of the matrix  $A$ . Indeed, we have  $\det A = \pm \det T$  according to the number of permutations carried out.

5. To obtain better numerical stability in the calculations on computers we should choose the pivot  $\tilde{a}_{kk}^k$  carefully. To avoid the propagation of rounding errors we must choose the largest pivot possible in absolute value. Even when the natural pivot  $a_{kk}^k$  is not zero, we permute to put in its place a larger pivot  $\tilde{a}_{kk}^k$ . We say that we make a partial pivot if we take the largest pivot possible in the  $k$ th column below the diagonal (as we have done in the proof above). We say that we make a total pivot if we take the largest pivot possible in the submatrix below the diagonal of size  $(n - k) \times (n - k)$  (in this case we permute row and column).

•

## LU factorization

The LU method consists of factorizing the matrix  $A$  into a product of two triangular matrices  $A = LU$ , where  $L$  is lower triangular and  $U$  is upper triangular. It is in fact the same algorithm as Gaussian elimination in the particular case where **we never pivot**. Once we have established the LU factorization of  $A$ , the solution of the linear system  $Ax = b$  is equivalent to the simple solution of two triangular systems  $Ly = b$  then  $Ux = y$ .

**Proposition 13.1.15** *Take a matrix  $A = (a_{ij})_{1 \leq i, j \leq n}$  of order  $n$  such that all the diagonal submatrices of order  $k$ , defined by*

$$\Delta^k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix}$$

*are invertible. There exists a unique pair of matrices  $(L, U)$ , with  $U$  upper triangular, and  $L$  lower triangular having a diagonal 1, such that*

$$A = LU.$$

**Remark 13.1.16** The hypothesis of proposition 13.1.15 is not unreasonable. Indeed, it is true if, for example,  $A$  is positive definite. If  $\Delta^k$  is not invertible, then there exists a nonzero vector  $x^k \in \text{Ker} \Delta^k$  and complementing it by zeros, we construct a nonzero vector  $x = (x^k, 0)$  which satisfies  $Ax \cdot x = 0$ , which contradicts the positive definite character of  $A$ .

•

**Proof.** Let us assume that during Gaussian elimination we do not need to make permutations to change the pivot, that is to say that all the natural pivots  $a_{kk}^k$  are nonzero. Then, with the notation of proposition 13.1.13 we have  $A^n = E^{n-1} \dots E^1 A$  with  $E^k$  defined by (13.6). We set  $U = A^n$  and  $L = (E^1)^{-1} \dots (E^{n-1})^{-1}$ . Then we have  $A = LU$  and it simply remains to verify that  $L$  is lower triangular. An easy calculation shows that  $(E^k)^{-1}$  is obtained starting from  $E^k$  by changing the sign of the elements under the diagonal, that is to say, setting  $l_{ik} = a_{i,k}^k / a_{k,k}^k$ , for  $k + 1 \leq i \leq n$ ,

we have

$$E^k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & -l_{k+1,k} & \ddots \\ & & & \vdots & \ddots \\ 0 & & -l_{n,k} & & 1 \end{pmatrix}, \quad (E^k)^{-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & +l_{k+1,k} & \ddots \\ & & & \vdots & \ddots \\ 0 & & +l_{n,k} & & 1 \end{pmatrix}.$$

Another calculation shows that  $L$  is lower triangular and that its  $k$ th column is the same as that of  $(E^k)^{-1}$

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n,1} & \dots & l_{n,n-1} & 1 \end{pmatrix}.$$

We must now verify that the pivots are not zero under the hypothesis made on the matrices  $\Delta^k$ . We verify this by induction. The first pivot  $a_{11}$  is nonzero as it is equal to  $\Delta^1$  which is invertible. We assume that all the pivots up to the order  $k-1$  are nonzero. Let us show that the new pivot  $a_{kk}^k$  is also nonzero. As the  $k-1$  first pivots are nonzero we could calculate the matrix  $A^k$ . We then write the equality  $(E^1)^{-1} \dots (E^{k-1})^{-1} A^k = A$  in the form of an equality between block matrices

$$\begin{pmatrix} L_{11}^k & 0 \\ 5L_{21}^k & I \end{pmatrix} \begin{pmatrix} U_{11}^k & A_{12}^k \\ A_{21}^k & A_{22}^k \end{pmatrix} = \begin{pmatrix} \Delta^k & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

with  $U_{11}^k$ ,  $L_{11}^k$ , and  $\Delta^k$  square blocks of size  $k$ , and  $A_{22}^k$ ,  $I$ , and  $A_{22}$  square blocks of size  $n-k$ . By applying the rules for multiplication block matrices, we obtain

$$L_{11}^k U_{11}^k = \Delta^k,$$

where  $U_{11}^k$  is an upper triangular matrix, and  $L_{11}^k$  a lower triangular matrix with '1s' on the diagonal. We deduce that the matrix  $U_{11}^k = (L_{11}^k)^{-1} \Delta^k$  is invertible as the product of invertible matrices. Its determinant is therefore nonzero. Or

$$\det U_{11}^k = \prod_{i=1}^k a_{ii}^k \neq 0.$$

Therefore the pivot  $a_{kk}^k$  at the step  $k$  is nonzero.

It only remains to verify the uniqueness. Take two LU decompositions of the matrix  $A = L_1 U_1 = L_2 U_2$ . We deduce that  $L_2^{-1} L_1 = U_2 U_1^{-1}$ , where the matrix



$L_2^{-1}L_1$  is lower triangular and  $U_2U_1^{-1}$  is upper triangular from lemma 13.1.17. They are therefore both diagonal, and as the diagonal of  $L_2^{-1}L_1$  is composed of 1s, we have  $L_2^{-1}L_1 = U_2U_1^{-1} = I$ .  $\square$

**Lemma 13.1.17** *Let  $T$  be a lower triangular matrix. Its inverse (if one exists) is also a lower triangular matrix and its diagonal elements are the inverses of the diagonal elements of  $T$ . Let  $T'$  be another lower triangular matrix. The product  $TT'$  is also lower triangular, and its diagonal elements are the product of diagonal elements of  $T$  and of  $T'$ .*

We leave to the reader the elementary proof of lemma 13.1.17.

**Practical calculation of LU factorization.** We calculate the  $LU$  factorization (if it exists) of a matrix  $A$  by identification with the product  $LU$ . By setting  $A = (a_{ij})_{1 \leq i, j \leq n}$ , and

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n,1} & \dots & l_{n,n-1} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} u_{1,1} & \dots & \dots & u_{1,n} \\ 0 & u_{2,2} & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & u_{n,n} \end{pmatrix},$$

as  $L$  is lower triangular and  $U$  upper triangular, for  $1 \leq i, j \leq n$  it becomes

$$a_{i,j} = \sum_{k=1}^n l_{i,k} u_{k,j} = \sum_{k=1}^{\min(i,j)} l_{i,k} u_{k,j}.$$

Identifying the columns of  $A_{ij}$  in increasing order, we deduce the columns of  $L$  and of  $U$ . Thus, after having calculated the  $(j-1)$  first columns of  $L$  and of  $U$  as a function of the  $(j-1)$  first columns of  $A$ , we read the  $j$ th column of  $A$

$$\begin{aligned} a_{i,j} &= \sum_{k=1}^i l_{i,k} u_{k,j} \Rightarrow u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k} u_{k,j} \quad \text{for } 1 \leq i \leq j, \\ a_{i,j} &= \sum_{k=1}^j l_{i,k} u_{k,j} \Rightarrow l_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} l_{i,k} u_{k,j}}{u_{jj}} \quad \text{for } j+1 \leq i \leq n. \end{aligned}$$

We therefore calculate the first  $j$  components of the  $j$ th column of  $U$  and the last  $n-j$  components of the  $j$ th column of  $L$  as a function of their  $(j-1)$  first columns. We divide by the pivot  $u_{jj}$  which must therefore be nonzero!

**Numerical algorithm.** We write the algorithm corresponding to the LU decomposition method in a computational pseudo-language. We have seen that we traverse the matrix  $A$  column by column: at the step  $k$  we calculate the  $k$ th column of  $L$ ,

then we place zeros under the diagonal of the  $k$ th column by making linear combinations of the  $k$ th row with each of the rows from  $k + 1$  to  $n$ . Since at the step  $k$  the first  $k$  rows and the first  $k - 1$  columns of the matrix are no longer modified, we can store in the same array which initially contained the matrix  $A$ , the matrices  $A^k$  and  $L^k = (E^1)^{-1} \cdots (E^{k-1})^{-1}$  (we only store the nontrivial elements of  $L^k$  instead of the zeros of  $A^k$  under the diagonal of its first  $k - 1$  columns). At the end, this array will contain the matrices  $L$  and  $U$  ( $L$  in its lower part—without the diagonal of 1s—and  $U$  in its upper part).

```

For  $k = 1, n - 1$        $\leftarrow$  step  $k$ 
  For  $i = k + 1, n$      $\leftarrow$  row  $i$ 
     $a_{ik} = a_{ik}/a_{kk}$      $\leftarrow$  new column of  $L$ 
    For  $j = k + 1, n$ 
       $a_{ij} = a_{ij} - a_{ik}a_{kj}$   $\leftarrow$  combination of rows  $i$  and  $k$ 
    End of the loop in  $j$ 
  End of the loop in  $i$ 
End of the loop in  $k$ 

```

**Operations count.** To measure the efficiency of the LU decomposition algorithm we count the number of operations needed to carry it out (which will be proportional to its time of execution on a computer). We do not calculate this number of operations exactly, and we are content with the first term of its asymptotic expansion when the dimension  $n$  is large. Moreover, for simplicity we only count the multiplications and divisions (and not the additions whose number is in general the same order of size).

- Elimination or LU decomposition: the number of operations  $N_{\text{op}}$  is

$$N_{\text{op}} = \sum_{j=1}^{n-1} \sum_{i=j+1}^n \left( 1 + \sum_{k=j+1}^n 1 \right),$$

which, to the first order, gives  $N_{\text{op}} \approx n^3/3$ .

- Substitution (or back substitution–forward substitution on the two triangular systems): the number of operations  $N_{\text{op}}$  is given by the formula

$$N_{\text{op}} = 2 \sum_{j=1}^n j,$$

which, to the first order, gives  $N_{\text{op}} \approx n^2$ .

In total the solution of a linear system  $Ax = b$  by LU factorization needs  $N_{\text{op}} \approx n^3/3$  operations as  $n^2$  is negligible compared with  $n^3$  when  $n$  is large.

**Remark 13.1.18** We also use the method of LU factorization to calculate the determinant and the inverse of a matrix. To obtain  $A^{-1}$ , we decompose  $A$  into LU factors

and we solve  $n$  linear systems with, as right-hand sides, the basis vectors  $(e_i)_{1 \leq i \leq n}$  (two substitutions per solution, but the basis vectors  $e_i$  have many zero components, which makes the forward substitution step with  $L$  cheap). The number of operations to calculate  $A^{-1}$  is

$$N_{\text{op}} \approx \frac{n^3}{3} + \sum_{j=1}^n \frac{j^2}{2} + n \left( \frac{n^2}{2} \right) \approx n^3.$$

To calculate the determinant of  $A$ , we decompose  $A$  into LU factors and we calculate the determinant of  $U$  (that of  $L$  is 1), for which we only need to multiply the diagonal elements of  $U$  together ( $n-1$  multiplications). The number of operations to calculate  $\det A$  is therefore  $N_{\text{op}} \approx n^3/3$ . •

### The Cholesky method

This is a method which is only applicable to real symmetric, positive definite matrices. It consists of factorizing a matrix  $A$  in the form  $A = BB^*$  where  $B$  is a lower triangular matrix (and  $B^*$  its adjoint or transpose).

**Proposition 13.1.19** *Let  $A$  be a real symmetric, positive definite matrix. There exists a unique real matrix  $B$  which is lower triangular, such that all its diagonal elements are positive, and which satisfies*

$$A = BB^*.$$

**Proof.** By application of proposition 13.1.15, there exists a unique pair of matrices  $(L, U)$  such that  $A = LU$  with  $U$  upper triangular and  $L$  lower triangular having a diagonal of 1s. We denote by  $D$  the diagonal matrix defined by  $D = \text{diag}(\sqrt{u_{ii}})$ . It is possible to take the square root of the elements of the diagonal of  $U$  as a block matrix multiplication argument shows that  $\prod_{i=1}^k u_{ii} = \det \Delta^k > 0$ , where  $\Delta^k$  is the diagonal submatrix of order  $k$  extracted from  $A$ , therefore each  $u_{ii}$  is strictly positive. We then set  $B = LD$  and  $C = D^{-1}U$  which verifies  $A = BC$ . As  $A = A^*$ , we deduce  $C(B^*)^{-1} = B^{-1}(C^*)$ . From lemma 13.1.17 the matrix  $C(B^*)^{-1}$  is upper triangular, while  $B^{-1}C^*$  is lower triangular. They are therefore both diagonal. Moreover, the diagonal elements of  $B$  and  $C$  are the same, therefore, the diagonal of  $B^{-1}C^*$  is only composed of 1s, that is to say,  $C(B^*)^{-1} = B^{-1}C^* = I$ , therefore  $C = B^*$ . To show the uniqueness of the Cholesky decomposition, we assume that there exist two factorizations  $A = B_1B_1^* = B_2B_2^*$ , from which  $B_2^{-1}B_1 = B_2^*(B_1^*)^{-1}$ . From lemma 13.1.17, we deduce that  $B_2^{-1}B_1 = D = \text{diag}(d_1, \dots, d_n)$ , and therefore that  $A = B_2B_2^* = B_2(DD^*)B_2^*$ . As  $B_2$  is invertible, it becomes  $D^2 = I$  therefore  $d_i = \pm 1$ . Now all the diagonal coefficients of a Cholesky decomposition are positive by hypothesis. Therefore  $d_i = 1$ , which implies  $B_1 = B_2$ . □

**Practical calculation of Cholesky factorization.** In practice, we calculate the Cholesky factor  $B$  by identification in the equality  $A = BB^*$ . Let  $A = (a_{ij})_{1 \leq i, j \leq n}$ ,

$B = (b_{ij})_{1 \leq i, j \leq n}$  with  $b_{ij} = 0$  if  $i < j$ . For  $1 \leq i, j \leq n$ , it becomes

$$a_{ij} = \sum_{k=1}^n b_{ik} b_{jk} = \sum_{k=1}^{\min(i, j)} b_{ik} b_{jk}.$$

By identifying the columns of  $A$  in increasing order (or its rows, which comes to the same thing as  $A$  is symmetric) we deduce the columns of  $B$ . Thus, after having calculated the first  $(j-1)$  columns of  $B$  as a function of  $(j-1)$  first columns of  $A$ , we read the  $j$ th column of  $A$  below the diagonal

$$\begin{aligned} a_{jj} &= \sum_{k=1}^j (b_{jk})^2 \quad \Rightarrow \quad b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} (b_{jk})^2} \\ a_{i,j} &= \sum_{k=1}^j b_{jk} b_{i,k} \quad \Rightarrow \quad b_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} b_{jk} b_{i,k}}{b_{jj}} \quad \text{for } j+1 \leq i \leq n. \end{aligned}$$

We therefore calculate the  $j$ th column of  $B$  as a function of its  $(j-1)$  first columns. Because of the preceding theorem, we are sure that, if  $A$  is symmetric positive definite, the terms under the square roots are strictly positive. Conversely, if  $A$  is not positive definite, we will find that  $a_{jj} - \sum_{k=1}^{j-1} (b_{jk})^2 \leq 0$  for a certain range  $j$ , which prevents the algorithm from terminating.

**Numerical algorithm.** The Cholesky algorithm can be written in a very compact fashion by using one array which originally contains  $A$  which is replaced step by step with the factor  $B$ . Let us remark that it is enough to store the lower half of  $A$  since  $A$  is symmetric.

```

For  $j = 1, n$ 
  For  $k = 1, j-1$ 
     $a_{jj} = a_{jj} - (b_{jk})^2$ 
  End of the loop in  $k$ 
   $a_{jj} = \sqrt{a_{jj}}$ 
  For  $i = j+1, n$ 
    For  $k = 1, j-1$ 
       $a_{ij} = a_{ij} - b_{jk} b_{ik}$ 
    End of the loop in  $k$ 
     $a_{ij} = a_{ij} / a_{jj}$ 
  End of the loop in  $i$ 
End of the loop in  $j$ 

```

**Operations count.** To measure the efficiency of the Cholesky method we count the number of operations (only the multiplications) necessary to carry it out. The number of square roots is  $n$  which is negligible in this operations count.

- Cholesky factorization: the number of operations  $N_{\text{op}}$  is

$$N_{\text{op}} = \sum_{j=1}^n \left( (j-1) + \sum_{i=j+1}^n j \right),$$

which, to the first order, gives  $N_{\text{op}} \approx n^3/6$ .

- Substitution: we must make a back substitution and a forward substitution on the triangular systems associated with  $B$  and  $B^*$ . The number of operations is to the first order  $N_{\text{op}} \approx n^2$ .

The Cholesky method is therefore approximately **twice as quick** as the Gaussian elimination method for a symmetric positive definite matrix.

### Banded matrices and sparse matrices

When a matrix has many zero coefficients, we say that it is **sparse**. If the nonzero elements are near to the diagonal, we say that the matrix has a **banded** structure. For these two types of matrices (which appear naturally in the finite element method as in most of the other methods), we can improve the operations count and the storage necessary to solve a linear system. This gain is very important in practice.

**Definition 13.1.20** *A matrix  $A \in \mathcal{M}_n(\mathbb{R})$  is called a banded matrix, with half-bandwidth  $p \in \mathbb{N}$  if its elements satisfy  $a_{i,j} = 0$  for  $|i - j| > p$ . The size of the band is then  $2p + 1$ .*

The interest in banded matrices comes from the following property.

**Exercise 13.1.4** Show that the LU and Cholesky factorizations conserve the banded structure of matrices.

**Remark 13.1.21** While the LU and Cholesky factorizations preserve the banded structure of matrices, this is not so for their sparse structure. In general, if  $A$  is sparse (even inside a band), the factors  $L$  and  $U$ , or  $B$  and  $B^*$  are ‘full’ (the opposite of sparse) in the interior of the same band. •

The following exercise allows us to quantify the gain that comes from using banded matrices.

**Exercise 13.1.5** Show that, for a banded matrix of order  $n$  and half-bandwidth  $p$ , the operations count of LU factorization is  $\mathcal{O}(np^2)$  and that of Cholesky factorization is  $\mathcal{O}(np^2/2)$ .

Let us pass to the case of sparse matrices. Let us explain first of all how to store these sparse matrices in the computer memory. As we only store the nonzero elements of the matrix, we obtain an appreciable saving in space. We present a storage method, called **Morse storage**, in a simple example, given that we only use it in practice for large matrices. Take therefore the matrix

$$A = \begin{pmatrix} 9 & 0 & -3 & 0 \\ 7 & -1 & 0 & 4 \\ 0 & 5 & 2 & 0 \\ 1 & 0 & -1 & 2 \end{pmatrix}.$$

The elements of  $A$  are stored, row by row, in an array with only one dimension STOCKA. We define an array DEBUTL which indicates the beginnings of the rows of  $A$  in STOCKA: more precisely STOCKA(DEBUTL( $i$ )) is the first nonzero element of the row  $i$ . We also need an array INDICC which indicates the column of each element stored in STOCKA: if  $a_{i,j}$  is stored in STOCKA( $k$ ), then INDICC( $k$ ) =  $j$ . The number of nonzero elements of  $A$  is equal to the size of the vector INDICC (or of the vector STOCKA). The number of rows of  $A$  is equal to the size of the vector DEBUTL. In our example, we have

STOCKA	INDICC	DEBUTL
9	1	1
-3	3	3
7	1	6
-1	2	8
4	4	
5	2	
2	3	
1	1	
-1	3	
2	4	

In general, the LU and Cholesky factorizations of a sparse matrix produce ‘full’ factors with few nonzero elements (there exist algorithms which minimize this ‘filling’ but they are less efficient than in the case of banded matrices). In practice, sparse matrices are often associated with the iterative methods that we shall see in Section 13.1.4. Indeed, matrix-vector products are easy to evaluate with this type of storage.

### Equivalence of operations and the Strassen algorithm

We remark that all these methods for the solution of linear systems have an operations count of the order of  $n^3$ , as does the calculation of the inverse (see remark 13.1.18) or the multiplication of two matrices of order  $n$ . It is not a coincidence as the following result shows, in a surprising way, that the inversion of a matrix has the same complexity as the multiplication of two matrices, even if the inversion seems, at first glance, to be a more complicated operation.

**Lemma 13.1.22** *The inversion of matrices and matrix multiplication have the same asymptotic complexity, that is to say that if there exists an algorithm for one of these operations such that its number of operations is bounded by  $\mathcal{O}(n^\alpha)$  with  $\alpha \geq 2$ , then we can construct an algorithm for the other operation whose number of operations is also bounded by  $\mathcal{O}(n^\alpha)$ .*

**Proof.** Let  $I(n)$  be the number of operations to calculate  $A^{-1}$  by a given algorithm, such that there exist  $C$  and  $\alpha \geq 2$  satisfying  $I(n) \leq Cn^\alpha$ . Let us show that there exists an algorithm to calculate the product  $AB$  whose number of operations  $P(n)$  is such that there exists  $C'$  for which  $P(n) \leq C'n^\alpha$  with the same exponent  $\alpha$ . We remark that

$$\begin{pmatrix} I & A & 0 \\ 0 & I & B \\ 0 & 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -A & AB \\ 0 & I & -B \\ 0 & 0 & I \end{pmatrix}.$$

Consequently, the product  $AB$  is obtained by inverting a matrix three times larger. Therefore,

$$P(n) \leq I(3n) \leq C3^\alpha n^\alpha.$$

Now let  $P(n)$  be the number of operations to calculate  $AB$  by a given algorithm, such that there exists  $C$  and  $\alpha$  satisfying  $P(n) \leq Cn^\alpha$ . Let us show that there exists an algorithm to calculate  $A^{-1}$  whose number of operations  $I(n)$  is such that there exists  $C'$  for which  $I(n) \leq C'n^\alpha$  with the same exponent  $\alpha$ . We remark that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B\Delta^{-1}CA^{-1} & -A^{-1}B\Delta^{-1} \\ -\Delta^{-1}CA^{-1} & \Delta^{-1} \end{pmatrix},$$

with  $\Delta = D - CA^{-1}B$  (sometimes called the Schur complement). We deduce

$$I(2n) \leq 2I(n) + 6P(n),$$

if we neglect the additions. Iterating this formula for  $n = 2^k$ , we obtain

$$I(2^k) \leq 2^k I(1) + 6 \sum_{i=0}^{k-1} 2^{k-i-1} P(2^i) \leq C \left( 2^k + \sum_{i=0}^{k-1} 2^{k-i-1+\alpha i} \right).$$

As  $\alpha \geq 2$ , we deduce

$$I(2^k) \leq C' 2^{\alpha k}.$$

If  $n \neq 2^k$  for all  $k$ , there exists  $k$  such that  $2^k < n < 2^{k+1}$ . We inscribe the matrix  $A$  in a larger matrix of size  $2^{k+1}$

$$\begin{pmatrix} A & 0 \\ 0 & I \end{pmatrix},$$

where  $I$  is the identity of order  $2^{k+1} - n$ . We obtain

$$I(n) \leq C'(2^{k+1})^\alpha \leq C' 2^\alpha n^\alpha,$$

which is the desired result.  $\square$

It was believed for a long time that the matrix multiplication of order  $n$  (and therefore their inversion) could not be done in less than  $n^3$  operations. But since a discovery by

Strassen in 1969, we know that this is not true. Indeed, Strassen developed an algorithm for matrix multiplication which needs far fewer operations for large  $n$ . He obtained, for this algorithm, an operations count

$$N_{\text{op}}(n) = \mathcal{O}(n^{\log_2 7}) \quad \text{with } \log_2 7 \sim 2, 81. \quad (13.7)$$

This result, itself surprising, has since been improved: we have found other algorithms, more and more complicated, for which the number of operations grows less quickly for large  $n$ , but we have not always found the best algorithm possible (that is to say that which leads to the smallest exponent  $\alpha$  such that  $N_{\text{op}}(n) = \mathcal{O}(n^\alpha)$ ). Obviously, for very large matrices, the gain in time obtained by these algorithms is appreciable (the Strassen algorithm has actually been used on supercomputers). Unfortunately, these algorithms often have numerical stability problems (they amplify the rounding errors) which limit their practical use.

The Strassen algorithm relies on the following lemma which appears benign!

**Lemma 13.1.23 (Strassen algorithm)** *The product of two matrices of order 2 can be made with 7 multiplications and 18 additions (instead of 8 multiplications and 4 additions by the usual rules).*

**Proof.** A simple calculation shows that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} m_1 + m_2 - m_4 + m_6 & m_4 + m_5 \\ m_6 + m_7 & m_2 - m_3 + m_5 - m_7 \end{pmatrix},$$

with

$$\begin{aligned} m_1 &= (b - d)(\gamma + \delta) & m_5 &= a(\beta - \delta) \\ m_2 &= (a + d)(\alpha + \delta) & m_6 &= d(\gamma - \alpha) \\ m_3 &= (a - c)(\alpha + \beta) & m_7 &= (c + d)\alpha \\ m_4 &= (a + b)\delta \end{aligned}$$

We count 7 multiplications and 18 additions. □

The crucial point in lemma 13.1.23 is that Strassen's multiplication rule is also valid if the matrix coefficients are in a noncommutative algebra. In particular, this is therefore true for block matrices.

Take then a matrix of size  $n = 2^k$ . We cut this matrix into four blocks of size  $2^{k-1}$ , and we apply the Strassen rule. If we count not only the multiplications but also the additions, the number of operations  $N_{\text{op}}(n)$  to make the product of two matrices satisfies

$$N_{\text{op}}(2^k) = 7N_{\text{op}}(2^{k-1}) + 18(2^{k-1})^2,$$

as the addition of two matrices of size  $n$  requires  $n^2$  additions. A simple induction gives

$$N_{\text{op}}(2^k) = 7^k N_{\text{op}}(1) + 18 \sum_{i=0}^{k-1} 7^i 4^{k-1-i} \leq 7^k (N_{\text{op}}(1) + 6).$$

We deduce easily that the optimal number of operations  $N_{\text{op}}(n)$  satisfies the bound (13.7).



### 13.1.4 Iterative methods

Iterative methods are particularly interesting for very large matrices or sparse matrices. Indeed, in this case direct methods can have a prohibitive calculation and storage cost (recall that LU or Cholesky factorization need of the order of  $n^3$  operations). Let us start with a very simple class of iterative methods.

**Definition 13.1.24** *Let  $A$  be an invertible matrix. We introduce a regular decomposition of  $A$  (sometime called a ‘splitting’), that is to say a pair of matrices  $(M, N)$  with  $M$  invertible (and easy to invert in practice) such that  $A = M - N$ . The iterative method based on the splitting  $(M, N)$  is defined by*

$$\begin{cases} x_0 \text{ given in } \mathbb{R}^n, \\ Mx_{k+1} = Nx_k + b \quad \forall k \geq 1. \end{cases} \quad (13.8)$$

If the sequence of approximate solutions  $x_k$  converges to a limit  $x$  when  $k$  tends to infinity, then, by passage to the limit in the induction (13.8), we obtain

$$(M - N)x = Ax = b.$$

Consequently, if the sequence of approximate solutions converges, its limit is the solution of the linear system.

From a practical point of view, we must know when to stop the iteration, that is to say at what moment  $x_k$  is sufficiently close to the unknown solution  $x$ . As we do not know  $x$ , we cannot decide to stop the calculation as soon as  $\|x - x_k\| \leq \epsilon$  where  $\epsilon$  is the desired precision. On the other hand, we know  $Ax$  (which is  $b$ ), and a stopping criterion frequently used is  $\|b - Ax_k\| \leq \epsilon$ . However, if the norm of  $A^{-1}$  is large this criterion can be misleading as

$$\|x - x_k\| \leq \|A^{-1}\| \|b - Ax_k\| \leq \epsilon \|A^{-1}\|$$

which will not be small.

**Definition 13.1.25** *We say that an iterative method is convergent if, whatever the choice of the initial vector  $x_0 \in \mathbb{R}^n$ , the sequence of approximate solutions  $x_k$  converges to the exact solution  $x$ .*

We start by giving a necessary and sufficient condition for convergence of an iterative method with the help of the spectral radius of the iteration matrix (see definition 13.1.5 for the idea of spectral radius).

**Lemma 13.1.26** *The iterative method defined by (13.8) converges if and only if the spectral radius of the iteration matrix  $M^{-1}N$  satisfies  $\rho(M^{-1}N) < 1$ .*

**Proof.** We define the error  $e_k = x_k - x$ . We have

$$e_k = (M^{-1}Nx_{k-1} + M^{-1}b) - (M^{-1}Nx + M^{-1}b) = M^{-1}Ne_{k-1} = (M^{-1}N)^k e_0.$$

By application of lemma 13.1.8 we deduce that  $e_k$  tends to 0, for any  $e_0$ , if and only if  $\rho(M^{-1}N) < 1$ .  $\square$

In practice the spectral radius of a matrix is difficult to calculate (we must calculate its eigenvalues). This is why we use other sufficient conditions of convergence, as indicated below.

**Lemma 13.1.27** *Let  $A$  be a Hermitian, positive definite matrix. Take a regular decomposition of  $A$  defined by  $A = M - N$  with  $M$  invertible. Then the matrix  $(M^* + N)$  is Hermitian. Moreover, if  $(M^* + N)$  is also positive definite, then*

$$\rho(M^{-1}N) < 1.$$

**Proof.** First, let us show that  $M^* + N$  is Hermitian:

$$(M^* + N)^* = M + N^* = (A + N) + N^* = A^* + N^* + N = M^* + N.$$

We define the vector norm  $|x|_A = \sqrt{\langle Ax, x \rangle}$  over  $\mathbb{R}^n$  (which is a norm as  $A$  is positive definite). We denote by  $\|\cdot\|$  the matrix norm subordinate to  $|\cdot|_A$ . We will show that  $\|M^{-1}N\| < 1$  which implies the desired result thanks to proposition 13.1.7. We calculate

$$\|M^{-1}N\|^2 = \max_{|v|_A=1} |M^{-1}Nv|_A^2.$$

Now, from lemma 13.1.2, there exists  $v$ , dependent on  $M^{-1}N$ , such that  $|v|_A = 1$  and

$$|M^{-1}Nv|_A^2 = \|M^{-1}N\|^2.$$

As  $N = M - A$ , we obtain, by setting  $w = M^{-1}Av$ ,

$$\begin{aligned} |M^{-1}Nv|_A^2 &= \langle AM^{-1}Nv, M^{-1}Nv \rangle \\ &= \langle AM^{-1}(M - A)v, M^{-1}(M - A)v \rangle \\ &= \langle (Av - AM^{-1}Av), (I - M^{-1}A)v \rangle \\ &= \langle Av, v \rangle - \langle AM^{-1}Av, v \rangle \\ &\quad + \langle AM^{-1}Av, M^{-1}Av \rangle - \langle Av, M^{-1}Av \rangle \\ &= 1 - \langle M^{-1}Av, MM^{-1}Av \rangle \\ &\quad + \langle AM^{-1}Av, M^{-1}Av \rangle - \langle MM^{-1}Av, M^{-1}Av \rangle \\ &= 1 - \langle w, Mw \rangle + \langle Aw, w \rangle - \langle Mw, w \rangle \\ &= 1 - \langle (M^* + N)w, w \rangle. \end{aligned}$$

Now  $\langle (M^* + N)w, w \rangle > 0$ , as  $(M^* + N)$  is positive definite, and  $w \neq 0$  as  $A$  and  $M$  are invertible. Therefore

$$\|M^{-1}N\|^2 = 1 - \langle (M^* + N)w, w \rangle < 1,$$

which finishes the proof.  $\square$

Since these iterative methods for the solution of linear systems are intended to be used on computers whose calculations are not exact but polluted by rounding errors, it is necessary to verify that these errors do not propagate to the point of destroying the convergence of the methods or, worse, make them converge to false solutions. Very happily this is not the case as the following result shows.

**Lemma 13.1.28** *Take a regular decomposition of  $A$  defined by  $A = M - N$  with  $M$  invertible. Take a right-hand side  $b \in \mathbb{R}^n$  and the solution  $x \in \mathbb{R}^n$  such that  $Ax = b$ . We assume that at each step  $k$  the iterative method has an error  $\epsilon_k \in \mathbb{R}^n$  in the sense that  $x_{k+1}$  is not exactly given by (13.8) but rather by*

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b + \epsilon_k.$$

*We assume that  $\rho(M^{-1}N) < 1$  and that there exists a vector norm and a positive constant  $\epsilon$  such that for all  $k \geq 0$*

$$\|\epsilon_k\| \leq \epsilon.$$

*Then there exists a constant  $K$ , which only depends on  $M^{-1}N$ , such that*

$$\limsup_{k \rightarrow +\infty} \|x_k - x\| \leq K\epsilon.$$

**Proof.** We again define the error  $e_k = x_k - x$  and we have now  $e_{k+1} = M^{-1}Ne_k + \epsilon_k$ . We deduce

$$e_k = (M^{-1}N)^k e_0 + \sum_{i=0}^{k-1} (M^{-1}N)^i \epsilon_{k-i-1}. \quad (13.9)$$

By application of proposition 13.1.7 there exists a subordinate matrix norm  $\|\cdot\|_s$  such that  $\|M^{-1}N\|_s < 1$  since  $\rho(M^{-1}N) < 1$ . We denote in the same way the associated vector norm. Now all the vector norms on  $\mathbb{R}^n$  are equivalent: there, therefore, exists a constant  $C \geq 1$ , which only depends on  $M^{-1}N$ , such that

$$C^{-1}\|y\| \leq \|y\|_s \leq C\|y\| \quad \forall y \in \mathbb{R}^n.$$

By bounding (13.9) above it becomes

$$\|e_k\|_s \leq \|M^{-1}N\|_s^k \|e_0\|_s + \sum_{i=0}^{k-1} \|M^{-1}N\|_s^i C\epsilon \leq \|M^{-1}N\|_s^k \|e_0\|_s + \frac{C\epsilon}{1 - \|M^{-1}N\|_s}$$

from which we obtain the result with  $K = C^2/(1 - \|M^{-1}N\|_s)$ . □

It is time to give the most classical examples of iterative methods based on a regular decomposition.

**Definition 13.1.29 (Jacobi method)** *Take  $A = (a_{ij})_{1 \leq i, j \leq n}$ . We denote by  $D = \text{diag}(a_{ii})$  the diagonal of  $A$ . We say Jacobi method for the iterative method associated with the decomposition*

$$M = D, \quad N = D - A.$$

This is the simplest of iterative methods. So that it is well defined, the diagonal matrix  $D$  must, of course, be invertible. By application of lemma 13.1.27, in the case where  $A$  is real symmetric, the Jacobi method converges if  $A$  and  $2D - A$  are positive definite.

**Definition 13.1.30 (Gauss–Seidel method)** Take  $A = (a_{ij})_{1 \leq i, j \leq n}$ . We decompose  $A$  in the form  $A = D - E - F$  where  $D = \text{diag}(a_{ii})$  is the diagonal,  $-E$  is the (strictly) lower triangular part, and  $-F$  is the (strictly) upper triangular part of  $A$ . We say the Gauss–Seidel method for the iterative method associated with the decomposition

$$M = D - E, \quad N = F.$$

So that the Gauss–Seidel method is well defined, the matrix  $D - E$  must be invertible, that is to say that  $D$  is invertible (the matrix  $(D - E)$  is easy to invert as it is triangular). By application of lemma 13.1.27, if  $A$  is real symmetric positive definite, the Gauss–Seidel method converges.

**Definition 13.1.31 (method of successive over-relaxation (SOR))** Take  $\omega \in \mathbb{R}^+$ . We say method of successive over-relaxation (or SOR), for the parameter  $\omega$ , for the iterative method associated with the decomposition

$$M = \frac{D}{\omega} - E, \quad N = \frac{1 - \omega}{\omega} D + F$$

So that the method of successive over-relaxation is well defined, we must again have the matrix  $D$  invertible. For  $\omega = 1$ , we recover the Gauss–Seidel method. If  $\omega < 1$ , it is in fact a method of under-relaxation, while  $\omega > 1$  corresponds to a method of over-relaxation. In general, there exists an optimal parameter  $\omega_{\text{opt}}$  which minimizes the spectral radius of the iteration matrix  $M^{-1}N$ , and therefore which maximizes the rate of convergence.

**Exercise 13.1.6** Let  $A$  be a Hermitian positive definite matrix. Show that for all  $\omega \in ]0, 2[$ , the method of successive over-relaxation converges.

**Exercise 13.1.7** Show that, for the method of successive over-relaxation, we always have

$$\rho(M^{-1}N) \geq |1 - \omega|, \quad \forall \omega \neq 0,$$

and therefore that it only converges if  $0 < \omega < 2$ .

**Definition 13.1.32 (gradient method)** Take a real parameter  $\alpha \neq 0$ . We say gradient method for the iterative method associated with the decomposition

$$M = \frac{1}{\alpha} I \quad \text{and} \quad N = \left( \frac{1}{\alpha} I - A \right).$$

The gradient method seems more primitive than the preceding methods, but it has an interpretation as a method of minimization of the function  $f(x) = \frac{1}{2}Ax \cdot x - b \cdot x$  which gives it wide applicability.

**Lemma 13.1.33** *Let  $A$  be a matrix with eigenvalues  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ . If  $\lambda_1 \leq 0 \leq \lambda_n$ , then the gradient method does not converge for any value of  $\alpha$ . If  $0 < \lambda_1 \leq \dots \leq \lambda_n$ , then the gradient method converges if and only if  $0 < \alpha < 2/\lambda_n$ , and the optimal parameter  $\alpha$ , which minimizes  $\rho(M^{-1}N)$ , is*

$$\alpha_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_n} \quad \text{and} \quad \min_{\alpha} \rho(M^{-1}N) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}.$$

**Remark 13.1.34** If  $\lambda_1 \leq \dots \leq \lambda_n < 0$ , then we have the symmetric result to the positive definite case by changing  $\alpha$  to  $-\alpha$ . For the optimal parameter  $\alpha_{\text{opt}}$ , the spectral radius of the iteration matrix is an increasing function of the ratio  $\lambda_n/\lambda_1$ . If  $A$  is self-adjoint, then this ratio is none other than the condition number  $\text{cond}_2(A)$  of the matrix  $A$ . Consequently, the more the matrix  $A$  is well conditioned, the better is the convergence of the gradient method. •

**Proof.** From lemma 13.1.26, we know that the gradient method is convergent if and only if  $\rho(M^{-1}N) < 1$ . Now  $M^{-1}N = (I - \alpha A)$ , therefore,

$$\rho(M^{-1}N) < 1 \Leftrightarrow |1 - \alpha\lambda_i| < 1 \Leftrightarrow -1 < 1 - \alpha\lambda_i < 1, \quad \forall i.$$

This implies that  $\alpha\lambda_i > 0$  for all  $1 \leq i \leq n$ . Consequently, all the eigenvalues of  $A$  must be nonzero and of the same sign as  $\alpha$ . The gradient method, therefore, does not converge if  $\lambda_1 \leq 0 \leq \lambda_n$ , no matter what  $\alpha$  is. If, on the contrary, we have  $0 < \lambda_1 \leq \dots \leq \lambda_n$ , then we deduce that we must have  $0 < \alpha < 2/\lambda_n$ . To calculate the optimal parameter  $\alpha_{\text{opt}}$ , we remark that the function  $\lambda \rightarrow |1 - \alpha\lambda|$  is decreasing on  $] -\infty, 1/\alpha]$  then increasing on  $[1/\alpha, +\infty[$ , therefore

$$\rho(M^{-1}N) = \max\{|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|\}.$$

Consequently, the minimum of the function  $\alpha \rightarrow \rho(M^{-1}N)$  is attained at the point of intersection  $\alpha_{\text{opt}} = 2/(\lambda_1 + \lambda_n)$  of these two lines.  $\square$

### 13.1.5 The conjugate gradient method

The conjugate gradient method is the iterative method of choice to solve linear systems whose matrix is real symmetric positive definite. It displays a spectacular improvement on the gradient method (above all if it is combined with a preconditioner, see definition 13.1.43). To construct the conjugate gradient method, we introduce the idea of a Krylov space.

**Definition 13.1.35** *Let  $r_0$  be a vector in  $\mathbb{R}^n$ . We say the Krylov space associated with the vector  $r_0$ , denoted by  $K_k$ , for the vector subspace of  $\mathbb{R}^n$  generated by the  $k+1$  vectors  $\{r_0, Ar_0, \dots, A^k r_0\}$ .*

The Krylov spaces  $(K_k)_{k \geq 0}$  form an increasing sequence of vector subspaces  $K_k \subset K_{k+1} \quad \forall k \geq 0$ . As  $K_k \subset \mathbb{R}^n$ , this sequence must become stationary at a certain point.

More precisely, we leave the reader to verify that there exists a critical dimension  $k_0$ , with  $0 \leq k_0 \leq n-1$ , such that

$$\begin{cases} \dim K_k = k+1 & \text{if } 0 \leq k \leq k_0, \\ \dim K_k = k_0+1 & \text{if } k_0 \leq k. \end{cases}$$

By taking  $r_0 = b - Ax_0$ , it is easy to see that, for the gradient method, the iterate  $x_k$  belongs to the affine space  $[x_0 + K_{k-1}]$  (defined as the set of vectors  $x$  such that  $(x - x_0) \in K_{k-1}$ ), which implies that the residual  $r_k = b - Ax_k$  belongs to the Krylov space  $K_k$  (associated with the residual initial  $r_0$ ).

To improve the gradient method, we decide ‘to choose better’  $x_k$  in the affine space  $[x_0 + K_{k-1}]$ . The conjugate gradient method consists, starting from an initial vector  $x_0 \in \mathbb{R}^n$  and from its residual  $r_0 = b - Ax_0$ , of constructing a sequence of vectors  $x_k \in [x_0 + K_{k-1}]$  such that  $r_k = b - Ax_k$  is orthogonal to the subspace  $K_{k-1}$ , for  $k \geq 1$ .

**Lemma 13.1.36** *Let  $A$  be a real symmetric positive definite matrix of order  $n$ . Take  $x_0 \in \mathbb{R}^n$ ,  $r_0 = b - Ax_0$ , and  $(K_k)_{k \geq 0}$  the sequence of Krylov spaces associated with  $r_0$ . The conjugate gradient method is defined, for  $k \geq 1$ , by*

$$x_k \in [x_0 + K_{k-1}] \quad \text{and} \quad r_k = b - Ax_k \perp K_{k-1}. \quad (13.10)$$

*For all  $k \geq 1$ , there exists a unique vector  $x_k$  given by (13.10). Moreover, this method converges to the solution of the linear system  $Ax = b$  in  $n$  iterations.*

**Remark 13.1.37** Lemma 13.1.36 shows that the conjugate gradient algorithm that we have conceived as an iterative method is in fact a direct method since it converges in a finite number of iterations (exactly  $k_0 + 1$  where  $k_0$  is the critical Krylov dimension). However, in practice we use it as an iterative method which converges ‘numerically’ in often less than  $k_0 + 1$  iterations. Intuitively, it is easy to see why the conjugate gradient improves on the simple gradient. Indeed, the residual  $r_k$  is orthogonal to an increasingly large subspace  $K_k$ . •

**Proof.** Let us show first of all that there exists a unique  $x_k$  which satisfies the hypotheses. As  $A$  is positive definite, we can define the scalar product  $\langle x, y \rangle_A = Ax \cdot y$  over  $\mathbb{R}^n$ . We look for  $x_k$  in the form  $x_k = x_0 + y_k$  with  $y_k \in K_{k-1}$ , and the condition of orthogonality of  $r_k$  becomes

$$\langle A^{-1}r_0 - y_k, y \rangle_A = 0 \quad \forall y \in K_{k-1},$$

which is none other than the characterization of  $y_k$  as the orthogonal projection of  $A^{-1}r_0$  on the subspace  $K_{k-1}$  (for the scalar product  $\langle \cdot, \cdot \rangle_A$ ). This proves the existence and uniqueness of  $x_k$ .

Let  $k_0$  be the critical dimension of the Krylov spaces, that is to say, for all  $k \geq k_0$ ,  $\dim K_k = k_0 + 1$ . In particular,  $AK_{k_0} \subset K_{k_0+1} = K_{k_0}$ , therefore  $r_{k_0+1} = b - Ax_{k_0+1} = r_0 - Ay_{k_0+1}$  belongs to  $K_{k_0}$  while being orthogonal. Consequently,  $r_{k_0+1} = 0$  and  $x_{k_0+1}$  is the exact solution of the linear system.  $\square$

**Remark 13.1.38** The conjugate gradient method has been presented as a procedure of orthogonalization with respect to the Krylov space. We see later that, equivalently, we can introduce it as a minimization problem. More precisely,  $x_k \in [x_0 + K_{k-1}]$  attains the minimum in  $[x_0 + K_{k-1}]$  of

$$f(x) = \frac{1}{2}Ax \cdot x - b \cdot x,$$

or equivalently the residual  $r_k = b - Ax_k$  minimizes, in  $K_k$ , the function

$$g(r) = \frac{1}{2}A^{-1}r \cdot r$$

with  $r = b - Ax$ . •

**Exercise 13.1.8** Let  $A$  be a symmetric positive definite matrix. Let  $(x_k)_{0 \leq k \leq n}$  be the sequence of approximate solutions obtained by the conjugate gradient method. We set  $r_k = b - Ax_k$  and  $d_k = x_{k+1} - x_k$ . Show that

(1) the Krylov space  $K_k$  is also equal to

$$K_k = [r_0, \dots, r_k] = [d_0, \dots, d_k],$$

(2) the sequence  $(r_k)_{0 \leq k \leq n-1}$  is orthogonal

$$r_k \cdot r_l = 0 \quad \text{for all } 0 \leq l < k \leq n-1,$$

(3) the sequence  $(d_k)_{0 \leq k \leq n-1}$  is conjugate with respect to  $A$

$$Ad_k \cdot d_l = 0 \quad \text{for all } 0 \leq l < k \leq n-1.$$

The definition we have just given of the conjugate gradient method is purely theoretical. Indeed, we have not indicated an algorithm to construct  $r_k$  orthogonal to  $K_{k-1}$ , nor commented on how we calculate  $x_k$  in practice. The following proposition gives particularly simple practical formulas to calculate these vectors.

**Proposition 13.1.39** Let  $A$  be a symmetric positive definite matrix, and  $x_0 \in \mathbb{R}^n$ . Let  $(x_k, r_k, p_k)$  be three sequences defined by the induction relations

$$p_0 = r_0 = b - Ax_0, \quad \text{and for } 0 \leq k \quad \begin{cases} x_{k+1} = x_k + \alpha_k p_k \\ r_{k+1} = r_k - \alpha_k A p_k \\ p_{k+1} = r_{k+1} + \beta_k p_k \end{cases} \quad (13.11)$$

with

$$\alpha_k = \frac{\|r_k\|^2}{Ap_k \cdot p_k} \quad \text{and} \quad \beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}.$$

Then,  $(x_k)_{0 \leq k \leq k_0+1}$  is the sequence of approximate solutions of the conjugate gradient method defined by (13.10).

**Proof.** It is easy to show by induction that the relations

$$r_0 = b - Ax_0 \quad \text{and} \quad \begin{cases} r_{k+1} = r_k - \alpha_k Ap_k \\ x_{k+1} = x_k + \alpha_k p_k \end{cases}$$

imply that the sequence  $r_k$  is that of the residual, that is,  $r_k = b - Ax_k$ . Another easy induction shows that the relations

$$r_0 = p_0 \quad \text{and} \quad \begin{cases} r_k = r_{k-1} - \alpha_{k-1} Ap_{k-1} \\ p_k = r_k + \beta_{k-1} p_{k-1} \end{cases}$$

imply that  $p_k$  and  $r_k$  belong to the Krylov space  $K_k$ , for all  $k \geq 0$ . We deduce, by the induction relation  $x_{k+1} = x_k + \alpha_k p_k$ , that  $x_{k+1}$  belongs to the affine space  $[x_0 + K_k]$ . To conclude, we must show that  $r_{k+1}$  is orthogonal to  $K_k$ . First, we shall show by induction that  $r_{k+1}$  is orthogonal to  $r_j$ , for all  $0 \leq j \leq k$ , and that  $p_{k+1}$  is conjugate to  $p_j$ , for all  $0 \leq j \leq k$ , that is to say,  $Ap_{k+1} \cdot p_j = 0$ . For the index 0 we have

$$r_1 \cdot r_0 = \|r_0\|^2 - \alpha_0 Ap_0 \cdot r_0 = 0$$

as  $p_0 = r_0$ , and

$$Ap_1 \cdot p_0 = (r_1 + \beta_0 p_0) \cdot Ap_0 = \alpha_0^{-1} (r_1 + \beta_0 r_0) \cdot (r_0 - r_1) = 0.$$

We assume that up to the index  $k$  we have

$$r_k \cdot r_j = 0 \text{ for } 0 \leq j \leq k-1 \quad \text{and} \quad Ap_k \cdot p_j = 0 \text{ for } 0 \leq j \leq k-1.$$

Let us show that this is again true for the index  $k+1$ . Because of the induction formula which gives  $x_{k+1}$  we have

$$r_{k+1} \cdot r_j = r_k \cdot r_j - \alpha_k Ap_k \cdot r_j,$$

and because of the relation  $r_j = p_j - \beta_{j-1} p_{j-1}$  we obtain

$$r_{k+1} \cdot r_j = r_k \cdot r_j - \alpha_k Ap_k \cdot p_j + \alpha_k \beta_{j-1} Ap_k \cdot p_{j-1}.$$

Because of the induction hypothesis, we deduce easily that  $r_{k+1} \cdot r_j = 0$  if  $j \leq k-1$ , while the formula for  $\alpha_k$  implies that  $r_{k+1} \cdot r_k = 0$ . On the other hand, the induction formula which gives  $p_{k+1}$  leads to

$$Ap_{k+1} \cdot p_j = p_{k+1} \cdot Ap_j = r_{k+1} \cdot Ap_j + \beta_k p_k \cdot Ap_j,$$

and as  $Ap_j = (r_j - r_{j+1})/\alpha_j$  we deduce

$$Ap_{k+1} \cdot p_j = \alpha_j^{-1} r_{k+1} \cdot (r_j - r_{j+1}) + \beta_k p_k \cdot Ap_j.$$

For  $j \leq k-1$ , the induction hypothesis and the orthogonality of  $r_{k+1}$  (that we have just obtained) proves that  $Ap_{k+1} \cdot p_j = 0$ . For  $j = k$ , we obtain  $Ap_{k+1} \cdot p_k = 0$  thanks to the formulas giving  $\alpha_k$  and  $\beta_k$ . This finishes this induction. As the family  $(r_k)_{0 \leq k \leq k_0}$  is orthogonal, it is linearly independent as long as  $r_k \neq 0$ . Now  $r_k \in K_k$ , which implies  $K_k = [r_0, \dots, r_k]$  as these two spaces have the same dimension. Consequently,  $r_{k+1}$  is orthogonal to  $K_k$ , and the sequence  $x_k$  is that of the conjugate gradient.  $\square$



**Remark 13.1.40** We say that the sequence  $(p_k)$  is **conjugate** with respect to  $A$  as it is orthogonal for the scalar product  $\langle x, y \rangle_A = Ax \cdot y$ . It is this property which gives the name to the method.

The reader might wonder how the formulas (13.11) have been ‘invented’. In fact, there exists a reciprocal to proposition 13.1.39. More precisely, exercise 13.1.8 shows that the sequence  $d_k = x_{k+1} - x_k$  is conjugate with respect to  $A$  and that the subspace generated by  $(d_0, \dots, d_k)$  coincides with  $K_k$ . We deduce therefore practical means of constructing the sequence  $d_k$ : we apply the procedure of Gram–Schmidt orthonormalization to  $(r_0, \dots, A^k r_0)$  for the scalar product  $\langle x, y \rangle_A$ . The result found is the sequence  $p_k$  which is necessarily colinear to  $d_k$  (we find that  $d_k = \alpha_k p_k$ ). Thanks to the symmetry of  $A$ , the Gram–Schmidt formulas which define  $p_k$  simplify considerably, and we are led (after some calculation) to the formulas (13.11). •

**Numerical algorithm.** In practice, when we apply the conjugate gradient algorithm, the formulas (13.11) of proposition 13.1.39 are programmed in the following way

$$\begin{array}{ll} \text{initialization} & \left\{ \begin{array}{l} \text{initial choice } x_0 \\ r_0 = p_0 = b - Ax_0 \end{array} \right. \\ \\ \text{iterations } k \geq 1 & \left\{ \begin{array}{l} \alpha_{k-1} = \|r_{k-1}\|^2 / Ap_{k-1} \cdot p_{k-1} \\ x_k = x_{k-1} + \alpha_{k-1} p_{k-1} \\ r_k = r_{k-1} - \alpha_{k-1} Ap_{k-1} \\ \beta_{k-1} = \|r_k\|^2 / \|r_{k-1}\|^2 \\ p_k = r_k + \beta_{k-1} p_{k-1} \end{array} \right. \end{array}$$

As soon as  $r_k = 0$ , the algorithm has converged, that is to say that  $x_k$  is the solution of the system  $Ax = b$ . We know that the convergence is attained in  $k_0 + 1$  iterations, where  $k_0 \leq n - 1$  is the critical dimension of the Krylov spaces (which we do not know *a priori*). However, in practice, the calculations on computers are often subject to rounding errors, and we do not find  $r_{k_0+1} = 0$  exactly. This is why, we introduce a ‘small’ parameter  $\epsilon$  (typically  $10^{-4}$  or  $10^{-8}$  according to the desired precision), and we decide that the algorithm has converged as soon as

$$\frac{\|r_k\|}{\|r_0\|} \leq \epsilon.$$

In addition, for large systems (for which  $n$  and  $k_0$  are ‘large’, of the order of  $10^4$ – $10^6$ ), the conjugate gradient method is used as an iterative method, that is to say that it converges, in the sense of the criterion above, in a number of iterations much less than  $k_0 + 1$  (cf. proposition 13.1.42).

**Remark 13.1.41**

1. In general, if we do not have any idea about the solution, we choose to initialize the conjugate gradient method by  $x_0 = 0$ . If we solve a sequence of problems a little different from each other, we can initialize  $x_0$  by the preceding solution.

2. At each iteration we only need to make one matrix-vector product, that is  $Ap_k$ , as  $r_k$  is calculated by the induction formula and not by the relation  $r_k = b - Ax_k$ .
3. To implement the conjugate gradient method, it is not necessary to store the matrix  $A$  in an array if we know how to calculate the matrix-vector product  $Ay$  for every vector  $y$ .
4. The conjugate gradient method is very efficient and heavily used. There are many variants or generalizations, particularly in the case of nonsymmetric positive definite matrices.

•

**Exercise 13.1.9** If we consider the conjugate gradient method as a direct method, show that in the most unfavourable case,  $k_0 = n - 1$ , the number of operations (multiplications only) to solve a linear system is  $N_{\text{op}} = n^3(1 + o(1))$ .

We have the following convergence result (see [8]).

**Proposition 13.1.42** *Let  $A$  be a real symmetric positive definite matrix. Let  $x$  be the exact solution of the system  $Ax = b$ . Let  $x_k$  be the sequence of approximate solutions of the conjugate gradient method. Then*

$$\|x_k - x\|_2 \leq 2\sqrt{\text{cond}_2(A)} \left( \frac{\sqrt{\text{cond}_2(A)} - 1}{\sqrt{\text{cond}_2(A)} + 1} \right)^k \|x_0 - x\|_2.$$

We recall that in the case of a real symmetric positive definite matrix, the condition number is given by the formula  $\text{cond}_2(A) = \lambda_n/\lambda_1$ , where  $\lambda_1, \lambda_n$  are respectively the smallest and the largest eigenvalue of  $A$ . Proposition 13.1.42 improves lemma 13.1.33 which says that the conjugate gradient method converges much faster than the gradient method.

We deduce from this result three important consequences. First, the conjugate gradient method functions well as an iterative method. Indeed, even if we do not make the  $n$  iterations required for convergence, we reduce the error between  $x$  and  $x_k$  as we iterate. On the other hand, the rate of convergence depends on the square root of the condition number of  $A$ , and not on the condition number itself as for the simple gradient method. The conjugate gradient method therefore converges much faster than the simple gradient (we say that the convergence is quadratic instead of being linear). Finally, the convergence will be all the more rapid as  $\text{cond}_2(A)$  is close to 1, that is to say that  $A$  is well conditioned.

## Preconditioning

As the rate of convergence of the conjugate gradient method depends of the condition number of the matrix  $A$ , the idea of the preconditioner is to premultiply the linear system  $Ax = b$  by a matrix  $C^{-1}$  such that the condition number of  $(C^{-1}A)$  is smaller than that of  $A$ . In practice we choose a matrix  $C$  ‘close’ to  $A$  but more easy to invert.

**Definition 13.1.43** We will solve the linear system  $Ax = b$ . We say preconditioner of  $A$  for a matrix  $C$  (easy to invert) such that  $\text{cond}_2(C^{-1}A)$  is smaller than  $\text{cond}_2(A)$ . We call the equivalent system the preconditioned system  $C^{-1}Ax = C^{-1}b$ .

In general, the matrix  $C^{-1}A$  is no longer symmetric, which poses a problem to apply the conjugate gradient method. This is why we often apply a ‘symmetric’ preconditioner which we describe now. Let us assume (for simplicity) that  $C$  is a symmetric positive definite matrix which has a Cholesky decomposition  $C = BB^*$ . We replace the original system  $Ax = b$  by the equivalent system

$$\tilde{A}\tilde{x} = \tilde{b}, \quad \text{with} \quad \tilde{A} = B^{-1}AB^{-*}, \quad \tilde{b} = B^{-1}b, \quad \text{and} \quad \tilde{x} = B^*x. \quad (13.12)$$

The matrix  $\tilde{A}$  being symmetric positive definite, we can use the conjugate gradient algorithm to solve this system. Nevertheless, we often do not know the Cholesky factorization of  $C$  (or we do not want to calculate it on grounds of cost). There then exists an astute way to transform the conjugate gradient algorithm for the system (13.12) into an algorithm where only  $C$  appears (and not the factor  $B$ ).

**Exercise 13.1.10** We denote by a tilde  $\tilde{\cdot}$  all the quantities associated with the conjugate gradient algorithm applied to the linear system (13.12). Take  $x_k = B^{-*}\tilde{x}_k$ ,  $r_k = B\tilde{r}_k = b - Ax_k$ , and  $p_k = B^{-*}\tilde{p}_k$ . Show that the conjugate gradient algorithm for (13.12) can also be written in the form

$$\begin{array}{ll} \text{initialization} & \left\{ \begin{array}{l} \text{initial choice } x_0 \\ r_0 = b - Ax_0 \\ p_0 = z_0 = C^{-1}r_0 \end{array} \right. \\ \\ \text{iterations } k \geq 1 & \left\{ \begin{array}{l} \alpha_{k-1} = z_{k-1} \cdot r_{k-1} / Ap_{k-1} \cdot p_{k-1} \\ x_k = x_{k-1} + \alpha_{k-1}p_{k-1} \\ r_k = r_{k-1} - \alpha_{k-1}Ap_{k-1} \\ z_k = C^{-1}r_k \\ \beta_{k-1} = z_k \cdot r_k / z_{k-1} \cdot r_{k-1} \\ p_k = z_k + \beta_{k-1}p_{k-1} \end{array} \right. \end{array}$$

The technique of preconditioning is very effective and essential in practice for rapid convergence. We indicate three possible choices of  $C$  from the most simple to the most complicated. The simplest preconditioner is the ‘diagonal preconditioner’: it consists of taking  $C = \text{diag}(A)$ . It is unfortunately not very effective, and we often prefer the ‘SSOR preconditioner’ (for symmetric successive over-relaxation). Denoting by  $D = \text{diag}(A)$  the diagonal of a symmetric matrix  $A$  and  $-E$  its strictly lower part such that  $A = D - E - E^*$ , for  $\omega \in ]0, 2[$ , we set

$$C_\omega = \frac{\omega}{2-\omega} \left( \frac{D}{\omega} - E \right) D^{-1} \left( \frac{D}{\omega} - E^* \right).$$

We verify that, if  $A$  is positive definite, then  $C$  is also. The system  $Cz = r$  is easy to solve as  $C$  is already in a factorised form as the product of triangular matrices. The name of this preconditioner comes from the fact that to invert  $C$  reduces to making two successive iterations of the successive over-relaxation (SOR) iterative method, with two symmetric iteration matrices.

**Exercise 13.1.11** Let  $A$  be the matrix of order  $n$  coming from the discretization of the Laplacian in  $N = 1$  dimension with a constant space step  $h = 1/(n + 1)$

$$A = h^{-1} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}.$$

Show that for the optimal value

$$\omega_{\text{opt}} = \frac{2}{1 + 2 \sin(\pi/2n)} \simeq 2 \left(1 - \frac{\pi}{n}\right)$$

the condition number of the matrix  $C_{\omega}^{-1}A$  is bounded above by

$$\text{cond}_2(C_{\omega}^{-1}A) \leq \frac{1}{2} + \frac{1}{2 \sin(\pi/2n)},$$

and therefore that, for  $n$  large, we gain an order in  $n$  in the rate of convergence.

A last example is the ‘incomplete Cholesky factorization preconditioner’. The matrix  $C$  is sought in the form  $BB^*$  where  $B$  is the ‘incomplete’ factor of the Cholesky factorization of  $A$  (see proposition 13.1.19). This lower triangular matrix  $B$  is obtained by applying the Cholesky factorization algorithm to  $A$  by forcing the equality  $b_{ij} = 0$  if  $a_{ij} = 0$ . This modification of the algorithm assures us, on the one hand, that the factor  $B$  will be as sparse as the matrix  $A$ , and, on the other hand, that the calculation of this incomplete factor will be much less expensive (in time of calculation) than the calculation of the exact factor if  $A$  is sparse (which is the case for finite element discretization matrices). The incomplete Cholesky factorization preconditioner is often the most effective preconditioner in practice.

## 13.2 Calculation of eigenvalues and eigenvectors

In this section we explain how to calculate the eigenvalues and the eigenvectors of a real symmetric matrix. A typical example is the stiffness matrix which comes from the approximation by finite elements of a partial differential equation. In this case,

its eigenvalues and eigenvectors are approximations of the eigenmodes of the physical model (see (7.23)).

Since the eigenvalues of a matrix  $A$  are the roots of its characteristic polynomial  $\det(A - \lambda I)$ . To calculate its eigenvalues, we could naively think that it is ‘sufficient’ to factorize its characteristic polynomial. It is nothing of the kind: we have known since Galois and Abel that we cannot calculate by elementary operations (addition, multiplication, extraction of roots) the roots of an arbitrary polynomial of degree greater than or equal to 5. To be convinced, we can notice that any polynomial of degree  $n$ ,

$$P(\lambda) = (-1)^n (\lambda^n + a_1 \lambda^{n-1} + a_2 \lambda^{n-2} + \cdots + a_{n-1} \lambda + a_n),$$

is the characteristic polynomial (expanded with respect to the last column) of the matrix

$$A = \begin{pmatrix} -a_1 & -a_2 & \cdots & \cdots & -a_n \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}.$$

Consequently, there cannot exist direct methods (that is to say which gives the result in a finite number of operations) for the calculation of eigenvalues! There therefore only exist iterative methods to calculate eigenvalues (and eigenvectors). We find that the practical calculation of eigenvalues and eigenvectors of a matrix is a much more difficult task than the solution of a linear system. Very fortunately, the case of real symmetric matrices (to which we limit ourselves as it is enough for our applications) is more simple than the case of non-self-adjoint matrices.

We indicate three typical methods (there are other, possibly more efficient but more complicated). The power method is the simplest but limited in its applicability. The Givens–Householder method allows us calculate one (or several) eigenvalue without having to calculate all the eigenvalues. Finally, the Lanczos method, which ‘resembles’ the conjugate gradient method, is at the foundation of many recent developments which leads to the most efficient methods for large sparse matrices.

### 13.2.1 The power method

A very simple method to calculate the largest or the smallest (in modulus) eigenvalue of a matrix and an associated eigenvector is the power method. A limitation of the method is that the extreme eigenvalue that we calculate must be simple (or of multiplicity equal to 1, that is to say that the dimension of the corresponding eigensubspace is 1). Let  $A$  be a real symmetric matrix of order  $n$ , with eigenvalues  $(\lambda_1, \dots, \lambda_n)$  with  $\lambda_n > |\lambda_i|$  for all  $1 \leq i \leq n-1$ . The power method to calculate the largest eigenvalue  $\lambda_n$  is defined by the algorithm below.

1. **Initialization:**  $x_0 \in \mathbb{R}^n$  such that  $\|x_0\| = 1$ .

2. Iterations: for  $k \geq 1$

1.  $y_k = Ax_{k-1}$
2.  $x_k = y_k / \|y_k\|$
3. convergence test: if  $\|x_k - x_{k-1}\| \leq \varepsilon$ , we stop.

In the convergence test  $\varepsilon$  is a small real number, typically equal to  $10^{-6}$ . If  $\delta_k = x_k - x_{k-1}$  is small, then  $x_k$  is an approximate eigenvector of  $A$  with approximate eigenvalue  $\|y_k\|$  as  $Ax_k - \|y_k\|x_k = A\delta_k$ .

**Proposition 13.2.1** *We assume that the matrix  $A$  is real symmetric, with eigenvalues  $(\lambda_1, \dots, \lambda_n)$ , associated with an orthonormal basis of eigenvectors  $(e_1, \dots, e_n)$ , and that the eigenvalue of largest modulus  $\lambda_n$  is simple and positive, that is to say that  $|\lambda_1|, \dots, |\lambda_{n-1}| < \lambda_n$ . We assume also that the initial vector  $x_0$  is not orthogonal to  $e_n$ . Then the power method converges, that is to say*

$$\lim_{k \rightarrow +\infty} \|y_k\| = \lambda_n, \quad \lim_{k \rightarrow +\infty} x_k = x_\infty \quad \text{with } x_\infty = \pm e_n.$$

The rate of convergence is proportional to the ratio  $|\lambda_{n-1}|/|\lambda_n|$

$$\| \|y_k\| - \lambda_n \| \leq C \left| \frac{\lambda_{n-1}}{\lambda_n} \right|^{2k}, \quad \|x_k - x_\infty\| \leq C \left| \frac{\lambda_{n-1}}{\lambda_n} \right|^k.$$

**Remark 13.2.2** The convergence of the sequence of approximate eigenvalues  $\|y_k\|$  is more rapid than that of the approximate eigenvectors  $x_k$  (quadratic instead of linear). The power method also works for nonsymmetric matrices, but the convergence of  $\|y_k\|$  is only linear in this case. •

**Proof.** Let  $x_0 = \sum_{i=1}^n \beta_i e_i$  the initial vector, with  $\beta_n \neq 0$ . The vector  $x_k$  is proportional to  $A^k x_0 = \sum_{i=1}^n \beta_i \lambda_i^k e_i$ , from which it becomes

$$x_k = \frac{\beta_n e_n + \sum_{i=1}^{n-1} \beta_i (\lambda_i / \lambda_n)^k e_i}{\left( \beta_n^2 + \sum_{i=1}^{n-1} \beta_i^2 (\lambda_i / \lambda_n)^{2k} \right)^{1/2}}.$$

As  $|\lambda_i| < \lambda_n$  we deduce that  $x_k$  converges to  $\text{sign}(\beta_n) e_n$ . Likewise, we have

$$\|y_{k+1}\| = \lambda_n \frac{\left( \beta_n^2 + \sum_{i=1}^{n-1} \beta_i^2 (\lambda_i / \lambda_n)^{2(k+1)} \right)^{1/2}}{\left( \beta_n^2 + \sum_{i=1}^{n-1} \beta_i^2 (\lambda_i / \lambda_n)^{2k} \right)^{1/2}},$$

which converges to  $\lambda_n$ . □

In practice (and particularly for the calculation of eigenvalues from the discretization of an elliptic boundary value problem), we are above all interested in the **smallest** eigenvalue, in modulus, of  $A$ . We can adapt the preceding ideas, which gives the inverse power method whose algorithm is written below. We consider a real symmetric matrix  $A$  whose smallest eigenvalue in modulus is simple and strictly positive  $0 < \lambda_1 < |\lambda_i|$  for all  $2 \leq i \leq n$ .

1. **Initialization:**  $x_0 \in \mathbb{R}^n$  such that  $\|x_0\| = 1$ .
2. **Iterations:** for  $k \geq 1$ 
  1. **solve**  $Ay_k = x_{k-1}$
  2.  $x_k = y_k / \|y_k\|$
  3. **convergence test:** if  $\|x_k - x_{k-1}\| \leq \varepsilon$ , we stop.

If  $\delta_k = x_k - x_{k-1}$  is small, then  $x_{k-1}$  is an approximate eigenvector of the approximate eigenvalue  $1/\|y_k\|$  as  $Ax_{k-1} - x_{k-1}/\|y_k\| = -A\delta_k$ .

**Proposition 13.2.3** *We assume that the matrix  $A$  is real symmetric, with eigenvalues  $(\lambda_1, \dots, \lambda_n)$ , associated with an orthonormal basis of eigenvectors  $(e_1, \dots, e_n)$ , and that the eigenvalue of smallest modulus  $\lambda_1$  is simple and strictly positive, that is to say  $0 < \lambda_1 < |\lambda_2|, \dots, |\lambda_n|$ . We assume also that the initial vector  $x_0$  is not orthogonal to  $e_1$ . Then the inverse power method converges, that is to say*

$$\lim_{k \rightarrow +\infty} \frac{1}{\|y_k\|} = |\lambda_1|, \quad \lim_{k \rightarrow +\infty} x_k = x_\infty \quad \text{with } x_\infty = \pm e_1.$$

The rate of convergence is proportional to the ratio  $\lambda_1/|\lambda_2|$

$$\left| \|y_k\|^{-1} - \lambda_1 \right| \leq C \left| \frac{\lambda_1}{\lambda_2} \right|^{2k}, \quad \|x_k - x_\infty\| \leq C \left| \frac{\lambda_1}{\lambda_2} \right|^k.$$

The proof is similar to that of proposition 13.2.1 and we leave it to the reader as an exercise.

**Remark 13.2.4** The considerations of remark 13.2.2 also apply to the inverse power method. To accelerate the convergence, we can often translate the matrix  $A$  and replace it by  $A - \sigma I$  with  $\sigma$  an approximation of  $\lambda_1$ . •

## 13.2.2 The Givens–Householder method

The Givens–Householder method decomposes into two successive steps: first of all the Householder algorithm which reduces a symmetric matrix  $A$  to a tridiagonal matrix (this step is carried out in a finite number of operations), then the Givens bisection algorithm which gives (iteratively) the eigenvalues of a tridiagonal matrix.

**Lemma 13.2.5 (Householder)** *Let  $A$  be a real symmetric matrix of order  $n$ . There exist  $(n - 2)$  orthogonal matrices  $H_k$  such that*

$$T = (H_1 H_2 \cdots H_{n-2})^* A (H_1 H_2 \cdots H_{n-2})$$

*which are tridiagonal. Of course,  $A$  and  $T$  have the same eigenvalues.*

**Proof.** Starting from  $A$ , we construct a sequence of matrices  $(A_k)_{1 \leq k \leq n-1}$  such that  $A_1 = A$  and  $A_{k+1} = H_k^* A_k H_k$  with  $H_k$  an orthogonal matrix chosen so that  $A_k$  has the following block structure

$$A_k = \begin{pmatrix} T_k & E_k^* \\ E_k & M_k \end{pmatrix}$$

where  $T_k$  is a square tridiagonal matrix of size  $k$ ,  $M_k$  is a square matrix of size  $n - k$ , and  $E_k$  is a rectangular matrix with  $(n - k)$  rows and  $k$  columns whose last column, denoted  $a_k \in \mathbb{R}^{n-k}$  is nonzero

$$T_k = \begin{pmatrix} \times & \times & & \\ & \times & \ddots & \\ & & \ddots & \times \\ & & & \times & \times \end{pmatrix}, \quad \text{and} \quad E_k = \begin{pmatrix} 0 & \cdots & 0 & a_{k,1} \\ \vdots & & \vdots & a_{k,2} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & a_{k,n-k} \end{pmatrix}.$$

It is clear that therefore  $A_{n-1}$  will be tridiagonal. We remark that  $A$  is in this form for  $k = 1$ . Let the matrix  $H_k$  be defined by

$$H_k = \begin{pmatrix} I_k & 0 \\ 0 & \tilde{H}_k \end{pmatrix},$$

with  $I_k$  the identity matrix of order  $k$  and  $\tilde{H}_k$  the Householder matrix of order  $n - k$  defined by

$$\tilde{H}_k = I_{n-k} - 2 \frac{v_k (v_k)^*}{\|v_k\|^2}, \quad \text{with } v_k = a_k + \|a_k\| e_1, \quad (13.13)$$

where  $e_1$  is the first vector of the canonical basis of  $\mathbb{R}^{n-k}$ . Let us remark that  $\tilde{H}_k a_k = -\|a_k\| e_1$ , and that  $H_k$  is orthogonal and symmetric. Let us note that  $\tilde{H}_k$  is only well defined if  $v_k \neq 0$ , but if this is not the case then the  $k$ th column of  $A_k$  is already of the desired type, and it is sufficient to take  $H_k = I_n$ . A simple calculation shows that

$$A_{k+1} = H_k^* A_k H_k = \begin{pmatrix} T_k & (\tilde{H}_k E_k)^* \\ \tilde{H}_k E_k & \tilde{H}_k M_k \tilde{H}_k \end{pmatrix} \quad \text{with } \tilde{H}_k E_k = \begin{pmatrix} 0 & \cdots & 0 & -\|a_k\| \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{pmatrix},$$

therefore  $A_{k+1}$  is in the form desired.  $\square$



Let us now study the Givens bisection algorithm for a real symmetric tridiagonal matrix

$$A = \begin{pmatrix} b_1 & c_1 & & & 0 \\ c_1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ 0 & & c_{n-1} & b_n & \end{pmatrix},$$

where we assume, without loss of generality, that  $c_i \neq 0$  for  $1 \leq i \leq n-1$ . Indeed, if there exists an index  $i$  such that  $c_i = 0$ , then we see easily that

$$\det(A - \lambda I) = \det(A_i - \lambda I) \det(A_{n-i} - \lambda I)$$

where  $A_i$  and  $A_{n-i}$  are two matrices of the same type as  $A$  but of order  $i$  and  $n-i$ , respectively. Let us first of all give two technical lemmas.

**Lemma 13.2.6** *For  $1 \leq i \leq n$ , we define a matrix  $A_i$  of size  $i$  by*

$$A_i = \begin{pmatrix} b_1 & c_1 & & & 0 \\ c_1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ 0 & & c_{i-1} & b_i & \end{pmatrix}.$$

*Let  $p_i(\lambda) = \det(A_i - \lambda I)$  be its characteristic polynomial. The sequence  $p_i$  satisfies the induction formula*

$$p_i(\lambda) = (b_i - \lambda)p_{i-1}(\lambda) - c_{i-1}^2 p_{i-2}(\lambda) \quad \forall i \geq 2,$$

*with  $p_1(\lambda) = b_1 - \lambda$  and  $p_0(\lambda) = 1$ . Moreover, for all  $i \geq 1$ , the polynomial  $p_i$  has the following properties*

- (1)  $\lim_{\lambda \rightarrow -\infty} p_i(\lambda) = +\infty$ ,
- (2) if  $p_i(\lambda_0) = 0$ , then  $p_{i-1}(\lambda_0)p_{i+1}(\lambda_0) < 0$ ,
- (3)  $p_i$  has  $i$  real distinct roots which strictly separate the  $(i+1)$  roots of  $p_{i+1}$ .

**Proof.** By expanding  $\det(A_i - \lambda I)$  with respect to the last row, we obtain the desired induction formula. The first property is obvious from the definition of the characteristic polynomial. To prove the second, we remark in the induction formula that, if  $p_i(\lambda_0) = 0$ , then

$$p_{i+1}(\lambda_0) = -c_i^2 p_{i-1}(\lambda_0).$$

As  $c_i \neq 0$ , we have  $p_{i-1}(\lambda_0)p_{i+1}(\lambda_0) \leq 0$ . This inequality is in fact strict, as if  $p_{i-1}(\lambda_0) = p_{i+1}(\lambda_0) = 0$  the induction relation implies that  $p_k(\lambda_0) = 0$  for all  $0 \leq k \leq i+1$ , which is not possible since  $p_0(\lambda_0) = 1$ .

To prove the third property, we start by remarking that  $p_i(\lambda)$  has  $i$  real roots, denoted  $\lambda_1^i \leq \dots \leq \lambda_i^i$  since  $A_i$  is real symmetric. Let us show by induction that these  $i$  roots of  $p_i$  are distinct and are separated by those of  $p_{i-1}$ . First, this property is true for  $i = 2$ , as

$$p_2(\lambda) = (b_2 - \lambda)(b_1 - \lambda) - c_1^2$$

has two roots ( $\lambda_1^2, \lambda_2^2$ ) which enclose the only root  $\lambda_1^1 = b_1$  of  $p_1(\lambda)$ , that is,  $\lambda_1^2 < \lambda_1^1 < \lambda_2^2$ . We assume that  $p_i(\lambda)$  has  $i$  real distinct roots separated by those of  $p_{i-1}$ . Let us show that  $p_{i+1}$  has  $i + 1$  real distinct roots separated by those of  $p_i$ . We define a polynomial  $q_i$  of degree  $2i$  by

$$q_i(\lambda) = p_{i-1}(\lambda)p_{i+1}(\lambda).$$

We already know  $i - 1$  roots of  $q_i$  (those of  $p_{i-1}$ ) and there are  $i$  values of  $\lambda$  (the roots of  $p_i$ ) such that  $q_i(\lambda) < 0$ , that is to say

$$q_i(\lambda_k^{i-1}) = 0 \quad 1 \leq k \leq i - 1, \quad q_i(\lambda_k^i) < 0 \quad 1 \leq k \leq i,$$

with

$$\lambda_1^i < \lambda_1^{i-1} < \lambda_2^i < \dots < \lambda_{i-1}^{i-1} < \lambda_i^i.$$

Between  $\lambda_k^i$  and  $\lambda_{k+1}^i$ , either  $q_i$  is cancelled by another term  $\gamma_k \neq \lambda_k^{i-1}$  and we have therefore found a supplementary root of  $q_i$  therefore of  $p_{i+1}$ , or  $q_i$  is only zero at  $\lambda_k^{i-1}$ , but in this case this is at least a double root as its derivative  $q_i'$  must also be zero at  $\lambda_k^{i-1}$ . However,  $\lambda_k^{i-1}$  is a simple root of  $p_{i-1}$ , therefore  $\lambda_k^{i-1}$  is also a root of  $p_{i+1}$ . But, because of the induction relation, this proves that  $\lambda_k^{i-1}$  is a root for all the polynomials  $p_j$  with  $0 \leq j \leq i + 1$ , which is not possible as  $p_0 = 1$ . Consequently, we have just shown that between each pair  $\lambda_k^i, \lambda_{k+1}^i$  there exists another root  $\gamma_k \neq \lambda_k^{i-1}$  of the polynomial  $q_i$  therefore of  $p_{i+1}$ . In total, we have found  $(i - 1)$  distinct roots of  $p_{i+1}$  which enclose those of  $p_i$ . In addition,  $q_i(\lambda_1^i) < 0$  and  $q_i(\lambda_i^i) < 0$ , while

$$\lim_{\lambda \rightarrow \pm\infty} q_i(\lambda) = +\infty.$$

We deduce the existence of two supplementary distinct roots of  $q_i$ , therefore of  $p_{i+1}$ , which enclose those of  $p_i$ .  $\square$

**Lemma 13.2.7** *For all  $\mu \in \mathbb{R}$ , we define*

$$\text{sgn}p_i(\mu) = \begin{cases} \text{sign of } p_i(\mu) & \text{if } p_i(\mu) \neq 0, \\ \text{sign of } p_{i-1}(\mu) & \text{if } p_i(\mu) = 0. \end{cases}$$

*Let  $N(i, \mu)$  the number of changes of sign between consecutive elements of the ordered family  $E(i, \mu) = \{+1, \text{sgn}p_1(\mu), \text{sgn}p_2(\mu), \dots, \text{sgn}p_i(\mu)\}$ . Then,  $N(i, \mu)$  is the number of roots of  $p_i$  which are strictly less than  $\mu$ .*

**Proof.** We remark first of all that  $\text{sgn}p_i(\mu)$  is defined unambiguously since, if  $p_i(\mu) = 0$ , then  $p_{i-1}(\mu) \neq 0$  because of point 2 of lemma 13.2.6. We proceed by inductions on  $i$ . For  $i = 1$ , we verify the result

$$\begin{aligned} \mu \leq b_1 &\Rightarrow E(1, \mu) = \{+1, +1\} \Rightarrow N(1, \mu) = 0, \\ \mu > b_1 &\Rightarrow E(1, \mu) = \{+1, -1\} \Rightarrow N(1, \mu) = 1. \end{aligned}$$

We assume the result true up to the index  $i$ . Let  $(\lambda_k^i)_{1 \leq k \leq i}$  be the roots of  $p_i$  and  $(\lambda_k^{i+1})_{1 \leq k \leq i+1}$  those of  $p_{i+1}$ , arranged in increasing order. We have

$$\lambda_1^i < \cdots < \lambda_{N(i,\mu)}^i < \mu \leq \lambda_{N(i,\mu)+1}^i < \cdots < \lambda_i^i,$$

and

$$\lambda_{N(i,\mu)}^i < \lambda_{N(i,\mu)+1}^{i+1} < \lambda_{N(i,\mu)+1}^i,$$

from point 3 of lemma 13.2.6. There are three possible cases.

1. If  $\lambda_{N(i,\mu)}^i < \mu \leq \lambda_{N(i,\mu)+1}^{i+1}$ , we have  $\operatorname{sgn} p_{i+1}(\mu) = \operatorname{sgn} p_i(\mu)$ , therefore  $N(i+1, \mu) = N(i, \mu)$ .
2. If  $\lambda_{N(i,\mu)+1}^{i+1} < \mu < \lambda_{N(i,\mu)+1}^i$ , we have  $\operatorname{sgn} p_{i+1}(\mu) = -\operatorname{sgn} p_i(\mu)$ , therefore  $N(i+1, \mu) = N(i, \mu) + 1$ .
3. If  $\mu = \lambda_{N(i,\mu)+1}^i$ , we have  $\operatorname{sgn} p_i(\mu) = \operatorname{sgn} p_{i-1}(\mu) = -\operatorname{sgn} p_{i+1}(\mu)$ , therefore  $N(i+1, \mu) = N(i, \mu) + 1$ , because of the second point of lemma 13.2.6.

In all the cases  $N(i+1, \mu)$  is the number of roots of  $p_{i+1}$  strictly less than  $\mu$ .  $\square$

**The practical Givens algorithm.** We denote by  $\lambda_1 \leq \cdots \leq \lambda_n$  the eigenvalues of  $A$  arranged in increasing order. To calculate numerically the  $i$ th eigenvalue  $\lambda_i$ , we take an interval  $[a_0, b_0]$  in which we are sure that  $\lambda_i$  can be found (for example,  $-a_0 = b_0 = \|A\|_2$ ). We then calculate the number  $N(n, (a_0 + b_0)/2)$  defined in lemma 13.2.7 (the values of the sequence  $p_j((a_0 + b_0)/2)$ , for  $1 \leq j \leq n$ , are calculated by the induction formula of lemma 13.2.6). If we find that  $N(n, (a_0 + b_0)/2) \geq i$ , then we deduce that  $\lambda_i$  belongs to the interval  $[a_0, (a_0 + b_0)/2]$ . If on the contrary we find that  $N(n, (a_0 + b_0)/2) < i$ , then  $\lambda_i$  belongs to the other interval  $[(a_0 + b_0)/2, b_0]$ . In each case we have divided the initial interval which contains  $\lambda_i$  by two. Repeating this procedure of division of the interval containing  $\lambda_i$  approximates the exact value of  $\lambda_i$  with the desired precision.

### 13.2.3 The Lanczos method

The Lanczos method allows us to calculate the eigenvalues of a real symmetric matrix by using the idea of a Krylov space, already introduced for the conjugate gradient algorithm. This method and its numerous generalizations is very effective for large matrices. Here we shall give the principle of this method rather than the details of its numerical implementation.

In what follows, we denote by  $A$  a real symmetric matrix of order  $n$ ,  $r_0 \neq 0 \in \mathbb{R}^n$  a given nonzero vector, and  $K_k$  the associated Krylov space, generated by the vectors  $\{r_0, Ar_0, \dots, A^k r_0\}$ . Let us recall that there exists an integer  $k_0 \leq n-1$ , called the critical Krylov dimension, such that, if  $k \leq k_0$ , the family  $(r_0, Ar_0, \dots, A^k r_0)$  is linearly independent and  $\dim K_k = k+1$ , while if  $k > k_0$  we have  $K_k = K_{k_0}$ .

The Lanczos algorithm consists of constructing a sequence of vectors  $(v_j)_{1 \leq j \leq k_0+1}$  by the following induction formula, for  $2 \leq j \leq k_0+1$ ,

$$\hat{v}_j = Av_{j-1} - (Av_{j-1} \cdot v_{j-1})v_{j-1} - \|\hat{v}_{j-1}\|v_{j-2}, \quad v_j = \frac{\hat{v}_j}{\|\hat{v}_j\|}, \quad (13.14)$$

with  $v_0 = 0$  and  $v_1 = r_0/\|r_0\|$ . For every integer  $k \leq k_0 + 1$ , we define a matrix  $V_k$  of size  $n \times k$  whose columns are the vectors  $(v_1, \dots, v_k)$ , as well as a symmetric **tridiagonal** matrix  $T_k$  of size  $k \times k$  whose elements are

$$(T_k)_{i,i} = Av_i \cdot v_i, \quad (T_k)_{i,i+1} = (T_k)_{i+1,i} = \|\hat{v}_{i+1}\|, \quad (T_k)_{i,j} = 0 \quad \text{if } |i-j| \geq 2.$$

With this notation, the Lanczos induction satisfies some remarkable properties.

**Lemma 13.2.8** *The sequence  $(v_j)_{1 \leq j \leq k_0+1}$  is well defined by (13.14) as  $\|\hat{v}_j\| \neq 0$  for all  $1 \leq j \leq k_0 + 1$ , while  $\hat{v}_{k_0+2} = 0$ . For  $1 \leq k \leq k_0 + 1$ , the family  $(v_1, \dots, v_{k+1})$  coincides with the orthonormal basis of  $K_k$  constructed by the Gram-Schmidt procedure applied to the family  $(r_0, Ar_0, \dots, A^k r_0)$ . Moreover, for  $1 \leq k \leq k_0 + 1$ , we have*

$$AV_k = V_k T_k + \hat{v}_{k+1} e_k^*, \quad (13.15)$$

where  $e_k$  is the  $k$ th vector of the canonical basis of  $\mathbb{R}^k$ ,

$$V_k^* AV_k = T_k \quad \text{and} \quad V_k^* V_k = I_k, \quad (13.16)$$

where  $I_k$  is the identity matrix of size  $k \times k$ .

**Proof.** Let us forget for the moment the definition (13.14) of the sequence  $(v_j)_{1 \leq j \leq k_0+1}$  and replace it by the new definition (which we show is equivalent to (13.14))

$$\hat{v}_j = Av_{j-1} - \sum_{i=1}^{j-1} (Av_{j-1} \cdot v_i) v_i, \quad v_j = \frac{\hat{v}_j}{\|\hat{v}_j\|}, \quad j \geq 2, \quad (13.17)$$

with  $v_1 = r_0/\|r_0\|$ . Of course, (13.17) only has a meaning if  $\|\hat{v}_j\| \neq 0$ . If  $\|\hat{v}_j\| = 0$ , we say that the algorithm stops at the index  $j$ . By definition,  $v_j$  is orthogonal to  $v_i$  for  $1 \leq i \leq j-1$ . By induction, we easily verify that  $v_j \in K_{j-1}$ . As the sequence of Krylov spaces  $K_j$  is strictly increasing for  $j \leq k_0 + 1$ , we deduce that, as long as the algorithm do not stop, the vectors  $(v_1, \dots, v_j)$  form an orthonormal basis of  $K_{j-1}$ . Consequently,  $v_j$  being orthogonal to  $(v_1, \dots, v_{j-1})$  is also orthogonal to  $K_{j-2}$ . In particular, the family  $(v_1, \dots, v_j)$ , defined by (13.17), coincides with the orthonormal basis of  $K_{j-1}$  constructed by the Gram-Schmidt procedure applied to the family  $(r_0, Ar_0, \dots, A^{j-1} r_0)$ . This proves that the algorithm stops exactly at the critical Krylov dimension  $k_0$ , that is to say that  $\|\hat{v}_j\| \neq 0$  as long as  $j \leq k_0 + 1$  and  $\hat{v}_{k_0+2} = 0$ .

Let us now show that the definitions (13.14) and (13.17) of the sequence  $(v_j)$  are identical. As  $A$  is symmetric, we have

$$Av_{j-1} \cdot v_i = v_{j-1} \cdot Av_i = v_{j-1} \cdot \hat{v}_{i+1} + \sum_{k=1}^i (Av_i \cdot v_k)(v_{j-1} \cdot v_k).$$

Thanks to the orthonormality properties of  $(v_k)$ , we deduce that  $Av_{j-1} \cdot v_i = 0$  if  $1 \leq i \leq j-3$ , and that  $Av_{j-1} \cdot v_{j-2} = \|\hat{v}_{j-1}\|$ . Therefore the definitions (13.14) and (13.17) coincide.

Finally, the relation (13.15), taken column by column, is none other than a rewriting of (13.14) by eliminating  $\hat{v}_j$ . The property  $V_k^* V_k = I_k$  follows from the orthonormal character of the family  $(v_1, \dots, v_k)$ , while the relation  $V_k^* AV_k = T_k$  is obtained simply by multiplying (13.15) to the left by  $V_k^*$  as  $V_k^* \hat{v}_{k+1} = 0$ .  $\square$

**Remark 13.2.9** The Lanczos algorithm appears to be a method of reduction to tridiagonal form like the Householder algorithm seen above. Nevertheless, the Lanczos algorithm is not used in practice as a method of tridiagonalization as, for  $n$  large, the rounding errors destroy the orthogonality of the last vectors  $v_j$  with respect to the first. •

We shall now compare the eigenvalues and eigenvectors of the matrices  $A$  and  $T_{k_0+1}$  (which are not of the same size in general). We denote by  $\lambda_1 < \lambda_2 < \dots < \lambda_m$  the distinct eigenvalues of  $A$  (with  $1 \leq m \leq n$ ), and  $P_1, \dots, P_m$  the matrices of orthogonal projection over the corresponding eigensubspaces of  $A$ . We recall that

$$A = \sum_{i=1}^m \lambda_i P_i, \quad I = \sum_{i=1}^m P_i, \quad \text{and} \quad P_i P_j = 0 \quad \text{if } i \neq j. \quad (13.18)$$

**Lemma 13.2.10** *The eigenvalues of  $T_{k_0+1}$  are simple and are also eigenvalues of  $A$ . Conversely, if we assume that  $r_0$  satisfies  $P_i r_0 \neq 0$  for all  $1 \leq i \leq m$ , then all the eigenvalues of  $A$  are also eigenvalues of  $T_{k_0+1}$  and  $k_0 + 1 = m$ .*

**Remark 13.2.11** In the case where  $P_i r_0 \neq 0$  for all  $i$ , the matrices  $A$  and  $T_{k_0+1}$  have exactly the same eigenvalues, but with possibly different multiplicity since the eigenvalues of  $T_{k_0+1}$  are simple. We shall see in the proof of lemma 13.2.10 that there also exists a link between the eigenvectors of  $A$  and  $T_{k_0+1}$ . The condition demanded on  $r_0$  for the reciprocal of this lemma is necessary. Indeed, if  $r_0$  is an eigenvector of  $A$ , then  $k_0 = 0$  and the matrix  $T_{k_0+1}$  has a unique eigenvalue which is associated with  $r_0$ . •

**Proof.** Let  $\lambda$  and  $y \in \mathbb{R}^{k_0+1}$  be an eigenvalue and an eigenvector such that  $T_{k_0+1} y = \lambda y$ . As  $\hat{v}_{k_0+2} = 0$ , the relation (13.15) becomes, for  $k = k_0 + 1$ ,  $AV_{k_0+1} = V_{k_0+1}T_{k_0+1}$ , and therefore, by application to the vector  $y$ , we obtain  $A(V_{k_0+1}y) = \lambda(V_{k_0+1}y)$ . The vector  $V_{k_0+1}y$  is not zero as  $y \neq 0$  and the columns of  $V_{k_0+1}$  are linearly independent. Consequently,  $V_{k_0+1}y$  is an eigenvector of  $A$  associated with the eigenvalue  $\lambda$  which is therefore equal to one of the  $\lambda_i$ .

Conversely, we introduce the vector subspace  $E_m$  of  $\mathbb{R}^n$  generated by the vectors  $(P_1 r_0, \dots, P_m r_0)$ . If  $P_i r_0 \neq 0$  for all  $1 \leq i \leq m$ , these vectors are linearly independent as the projections  $P_i$  are pairwise orthogonal. Consequently, the dimension of  $E_m$  is exactly  $m$ . We shall show that in this case we have  $m = k_0 + 1$ . By (13.18) we have  $A^k r_0 = \sum_{i=1}^m \lambda_i^k P_i r_0$ , that is to say  $A^k r_0 \in E_m$ , therefore the Krylov spaces satisfy  $K_k \subset E_m$  for all  $k \geq 0$ . In particular, this implies that  $\dim K_{k_0} = k_0 + 1 \leq m$ . In addition, in the basis  $(P_1 r_0, \dots, P_m r_0)$  of  $E_m$ , the coordinates of the vector  $A^k r_0$  are  $(\lambda_1^k, \dots, \lambda_m^k)$ . In other words, the family  $(r_0, Ar_0, \dots, A^{m-1}r_0)$  of  $E_m$  is represented in the basis  $(P_1 r_0, \dots, P_m r_0)$  by the matrix  $M$  defined by

$$M = \begin{pmatrix} 1 & \lambda_1 & \lambda_1^2 & \dots & \lambda_1^{m-1} \\ \vdots & & & & \vdots \\ 1 & \lambda_m & \lambda_m^2 & \dots & \lambda_m^{m-1} \end{pmatrix}.$$

The matrix  $M$  is a Van Der Monde matrix of order  $m$  which is invertible as all the  $\lambda_i$  are distinct. Therefore the family  $(r_0, Ar_0, \dots, A^{m-1}r_0)$  is linearly independent since  $(P_1r_0, \dots, P_mr_0)$  is. This implies that  $\dim K_{m-1} = m$ , therefore  $m - 1 \leq k_0$ . From where, we finally deduce that  $m = k_0 + 1$  and  $E_m = K_{k_0}$ .

Thanks to the formula (13.18) we have  $A(P_i r_0) = \lambda_i(P_i r_0)$ . As  $P_i r_0$  is nonzero, it is an eigenvector of  $A$  associated with the eigenvalue  $\lambda_i$ . As  $E_m = K_{k_0}$  and the columns of  $V_{k_0+1}$  form a basis of  $K_{k_0}$ , we deduce that there exists a nonzero vector  $y_i \in \mathbb{R}^m$  such that  $P_i r_0 = V_{k_0+1} y_i$ . We multiply the first equality of (13.16) by  $y_i$  to obtain

$$T_{k_0+1} y_i = V_{k_0+1}^* A V_{k_0+1} y_i = V_{k_0+1}^* A P_i r_0 = \lambda_i V_{k_0+1}^* P_i r_0 = \lambda_i V_{k_0+1}^* V_{k_0+1} y_i = \lambda_i y_i,$$

in other words,  $y_i$  is an eigenvector of  $T_{k_0+1}$  for the eigenvalue  $\lambda_i$ . Which finishes the proof.  $\square$

The result of lemma 13.2.10 could lead us to believe that we must apply Lanczos induction up to the maximal iteration  $k_0 + 1$ , then calculate the eigenvalues of  $T_{k_0+1}$  in order to deduce the eigenvalues and the eigenvectors of  $A$ . This would make the Lanczos method comparable to that of Givens–Householder (in general  $k_0$  is of the order of  $n$ , which makes the operations count similar in the two cases). Moreover, applied like this the Lanczos method will be numerically unstable because of the loss of orthogonality of the vectors  $v_j$  caused by the inevitable rounding errors (see remark 13.2.9).

Very happily, the following result indicates that it is not necessary to make many iterations of the Lanczos induction to obtain eigenvalues of  $T_k$  which are good approximations to those of  $A$  (with  $k$  much smaller than  $k_0$  or  $n$ ).

**Proposition 13.2.12** *Take an integer  $1 \leq k \leq k_0 + 1$ . Let  $\lambda$  be an eigenvalue of  $T_k$  and  $y \in \mathbb{R}^k$  an associated nonzero eigenvector. There exists an eigenvalue  $\lambda_i$  of  $A$  such that*

$$|\lambda - \lambda_i| \leq \|\hat{v}_{k+1}\| \frac{|e_k \cdot y|}{\|y\|} \leq \|\hat{v}_{k+1}\|,$$

where  $e_k$  is the  $k$ th vector of the canonical basis of  $\mathbb{R}^k$ .

**Remark 13.2.13** The first conclusion of proposition 13.2.12 is that, if  $\|\hat{v}_{k+1}\|$  is small, then the eigenvalues of  $T_k$  are good approximations of some eigenvalues of  $A$ . The second conclusion is the more important in practice: if the last component of an eigenvector of  $T_k$  is small, then the corresponding eigenvalue is a good approximation to an eigenvalue of  $A$ .  $\bullet$

**Proof.** Take a nonzero eigenvector  $y \in \mathbb{R}^k$  such that  $T_k y = \lambda y$ . Multiplying (13.15) by  $y$  we obtain

$$A V_k y = V_k T_k y + (e_k \cdot y) \hat{v}_{k+1},$$

from where we deduce

$$A(V_k y) - \lambda(V_k y) = (e_k \cdot y) \hat{v}_{k+1}. \quad (13.19)$$

We then decompose  $V_k y$  into the basis of eigenvectors of  $A$

$$V_k y = \sum_{i=1}^m P_i(V_k y).$$

We take the scalar product of (13.19) with  $V_k y$ , and with the help of the relations (13.18) we have

$$\sum_{i=1}^m (\lambda_i - \lambda) |P_i(V_k y)|^2 = (e_k \cdot y) (\hat{v}_{k+1} \cdot V_k y). \quad (13.20)$$

By bounding below the right-hand side and bounding above the left by the Cauchy–Schwarz inequality, it becomes

$$\min_{1 \leq i \leq m} |\lambda_i - \lambda| \|V_k y\|^2 \leq \|y\| \|\hat{v}_{k+1}\| \|V_k y\|.$$

As the columns of  $V_k$  are orthonormal, we have  $\|V_k y\| = \|y\|$ , and by simplification we obtain

$$\min_{1 \leq i \leq m} |\lambda_i - \lambda| \leq \|\hat{v}_{k+1}\|.$$

This inequality can be improved if we do not apply Cauchy–Schwarz to the term  $\langle e_k, y \rangle$  in (13.20). In this case we find

$$\min_{1 \leq i \leq m} |\lambda_i - \lambda| \leq \|\hat{v}_{k+1}\| \frac{|e_k \cdot y|}{\|y\|},$$

which finishes the proof. □

# Bibliography

- [1] AHUJA R.K., MAGNANTI T.L., ORLIN J.B., *Network flows*, Prentice Hall, Upper Saddle River, New Jersey (1993).
- [2] AZENCOTT R. editor, *Simulated annealing. Parallelization techniques*, Wiley-Interscience Series in Discrete Mathematics. A Wiley-Interscience Publication. John Wiley & Sons, New York (1992).
- [3] BONNANS J., GILBERT J.-C., LEMARECHAL C., SAGASTIZABAL C., *Numerical optimization. Theoretical and practical aspects*, Translated and revised from the 1997 French original. Universitext, Springer-Verlag, Berlin (2003).
- [4] BONY J.-M., *Cours d'analyse. Théorie des distributions et analyse de Fourier*, Éditions de l'École Polytechnique, Palaiseau (2001).
- [5] BRENNER S., SCOTT R., *The mathematical theory of finite element methods*, Texts in Applied Mathematics 15, Springer-Verlag, New York (2002).
- [6] BREZIS H., *Analyse fonctionnelle*, Masson, Paris (1983).
- [7] CHVÁTAL V., *Linear programming*, Freeman and Co., New York (1983).
- [8] CIARLET P.G., *Introduction to numerical linear algebra and optimisation*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge (1989).
- [9] CIARLET P.G., *The finite element methods for elliptic problems*, North-Holland, Amsterdam (1978).
- [10] CIARLET P.G., LIONS J.-L., *Handbook of numerical analysis*, North-Holland, Amsterdam (1990).
- [11] COOK J., CUNNINGHAM W.H., PULLEYBANK W.R., SCHRIJVER A., *Combinatorial optimization*, John Wiley & Sons, New York (1998).
- [12] COURANT R., HILBERT R., *Methods of mathematical physics*, John Wiley & Sons, New York (1989).



- [13] DANAILA I., JOLY P., KABER S. M., POSTEL M., *An introduction to scientific computing*, Springer-Verlag, Berlin (2006).
- [14] DAUTRAY R., LIONS J.-L., *Mathematical analysis and numerical methods for science and technology*, Springer-Verlag, Berlin (1988).
- [15] DUVAUT G., LIONS J.-L., *Inequalities in mechanics and physics*, Grundlehren der Mathematischen Wissenschaften 219, Springer-Verlag, Berlin (1976).
- [16] EKELAND I., TEMAM R., *Convex analysis and variational problems*, Classics in Applied Mathematics 28, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1999).
- [17] ERN A., GUERMOND J.-L., *Theory and practice of finite elements*, Applied Mathematical Sciences 159, Springer-Verlag, New York (2004).
- [18] EVANS L., *Partial differential equations*, Graduate Studies in Mathematics 19, American Mathematical Society, Providence, RI (1998).
- [19] FLETCHER R., *Practical methods of optimization*, John Wiley & Sons, New York (2001).
- [20] GEORGE P.L., *Automatic mesh generation. Application to finite element methods*, John Wiley & Sons, Chichester; Masson, Paris (1991).
- [21] GIRAULT V., RAVIART P.-A., *Finite element methods for Navier-Stokes equations: theory and algorithms*, Springer-Verlag, Berlin (1986).
- [22] GLOVER F., LAGUNA M., *Tabu search*, Kluwer, Boston (1997).
- [23] GODLEWSKI E., RAVIART P.-A., *Numerical approximation of hyperbolic systems of conservation laws*, Applied Mathematical Sciences 118, Springer-Verlag, New York (1996).
- [24] LAX P., *Linear algebra*. John Wiley & Sons, New York (1997).
- [25] LAX P., *Functional analysis*, John Wiley & Sons, New York (2002).
- [26] LEMAITRE J., CHABOCHE J.-L., *Mechanics of Solid Materials*, Cambridge University Press, Cambridge (1990).
- [27] LIONS J.-L., *Quelques méthodes de résolution des problèmes aux limites non-linéaires*, Dunod, Paris (1969).
- [28] LIONS J.-L., MAGENES E., *Non-homogeneous boundary value problems and applications*, Die Grundlehren der mathematischen Wissenschaften 181, Springer-Verlag, New York (1972).
- [29] LUCQUIN B., PIRONNEAU O., *Introduction to scientific computing*, John Wiley & Sons, Chichester (1998).

- [30] NOCEDAL J., WRIGHT S., *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, Springer-Verlag, New York (2006).
- [31] PADBERG M., *Linear optimization and extensions*, Springer-Verlag, Berlin (1999).
- [32] PIRONNEAU O., *Finite element methods for fluids*, John Wiley & Sons, Chichester; Masson, Paris (1989).
- [33] QUARTERONI A., SALERI F., *Scientific computing with MATLAB*, Texts in Computational Science and Engineering 2, Springer-Verlag, Berlin (2003).
- [34] RAVIART P.-A., THOMAS J.-M., *Introduction à l'analyse numérique des équations aux dérivées partielles*, Masson, Paris (1983).
- [35] RUDIN W., *Real and complex analysis*, Mc Graw Hill, New York (1966).
- [36] SALENÇON J., *Handbook of Continuum Mechanics*, Springer-Verlag, Berlin (2001).
- [37] SCHRIJVER A., *Theory of linear and integer programming*, John Wiley & Sons, New York (1986).
- [38] SCHWARTZ L., *Mathematics for the physical sciences*, Hermann, Paris (1966).
- [39] TEMAM R., *Navier-Stokes equations. Theory and numerical analysis*, AMS Chelsea Publishing, Providence, RI (2001).

*This page intentionally left blank*

# Index

- $N$ -rectangle 191
- $N$ -simplex 171
- $\alpha$ -convexity 290
  
- accuracy 35
- active constraint 312
- adjoint 209
- adjoint state 327
- admissible directions 306
- admissible solutions 350
- affectation 370
- alternating directions 50
- arcs 368
  
- Banach space 81
- banded matrix 420
- basic solution 351
- basis 351
- Bellman equation 372
- boundary value problem 25, 67
- bounded support 68, 87
- Bramble–Hilbert lemma 189
  
- calculus of variations 283
- capacity 368
- Cauchy problem 26, 67
- centred 15
- CFL condition 19, 24, 40, 47, 267, 272
- Cholesky factorization 418
- circuit 372
- coercive 74
- compact 209, 210
- compact support 80
  
- complementary energy 324
- complementary slack 361
- complete 399
- complete graph 384
- complexity 396, 421
- condition number 409
- conjugate gradient 335, 429
- conjugate gradient method 428
- connected 381
- connected components 381
- consistency 35
- constraint qualifications 312
- control 284, 330
- convex 289, 399
- convex envelope 362
- cost 368
- cost or objective function 284
- Céa’s lemma 151
  
- degrees of freedom 176
- differentiability in the sense of Fréchet 298
- differentiability in the sense of Gâteaux 300
- diffusion equation 2, 5
- diffusive 57, 274
- direct method 406
- Dirichlet boundary condition 4
- discrete norm 166
- dispersive 57
- distribution 105
- divergence 3
- domain of a convex function 391

- domain of dependence 10, 263
- dual 358, 402
- dual problem 320
- dynamic programming 372
- edges 381
- eigenfunction 214
- eigenmode 221
- eigenvalue 209
- eigenvector 209
- elasticity 12, 136
- elliptic 28, 74
- energy equality 71, 242
- energy estimate 113, 119, 141, 237
- energy space 84, 236, 242, 246, 251
- epigraph 290
- equipartition of energy 263
- equivalent equation 55
- Euler inequality 303
- explicit 16
- extremal point 362
- Farkas lemma 313
- finite differences 14
- finite velocity of propagation 8, 10
- flow 368
- Ford–Bellman algorithm 377
- forest 381
- Fourier boundary condition 4
- function value 372
- Gauss–Seidel method 427
- Gaussian elimination 412
- Givens–Householder method 438
- global minimum 285
- gradient 3, 427
- gradient algorithm with fixed step 335
- gradient algorithm with optimal step 333
- gradient method 427
- greedy algorithm 380
- Green’s formula 68, 90
- Green’s function 260
- heat flow equation 2
- Hermite finite elements 169
- Hilbert space 399
- Hilbertian basis 400
- hyperbolic 28
- hyperplane of support 362, 403
- implicit 16
- infimum 285
- infinite at infinity 285
- initial condition 3
- integer envelope 364
- integer point 362
- integer polyhedron 364
- interpolation 159, 186
- invariance under scale change 8
- irreversible 8
- iteration matrix 38
- iterative method 406
- Jacobi method 426
- Kirchoff law 368
- Kruskal algorithm 381
- Kuhn–Tucker theorem 319
- Lagrange finite elements 176
- Lagrange multiplier 305
- Lagrangian 310, 317
- Lagrangian relaxation 391
- Lamé 12
- Lanczos method 442
- Laplacian 3
- lattice 173, 192
- Lax–Milgram theorem 74
- linear form 402
- linear programming 348
- linear scheme 37
- linear system 405
- local minimum 285
- LU factorization 414
- mass matrix 225
- matrix assembly 178
- maximal solution 376

- maximum principle 8, 20, 38, 127, 165, 255, 262
- mesh 14, 153, 171, 191
- method of successive over-relaxation 427
- minimax principle 216
- minimizing sequence 285
- Minkowski theorem 362
- modelling 2, 4, 7, 14, 134, 136, 144, 255, 257
- multi-index 96
- multilevel schemes 33, 44
  
- Navier–Stokes 144
- Neumann boundary condition 4
- Newton’s method 342
- node 172
- nodes 368
- nonoriented graph 381
- NP-complete (problems) 397
- numerical convergence 168
- numerical diffusion 57
- numerical integration 157
  
- one-sided 16
- operational research 347
- operations research 278
- operator 402
- optimal command 326
- optimal solution 350
- OR 347
- order of a PDE 28
- order of a scheme 35
- oriented graph 368
- orthogonal projection 399
- outward normal 68
  
- parabolic 28
- partial differential equations 1
- path 372
- penalization 358
- penalization 341
- periodic boundary conditions 39
- plate equation 14, 170
- Poincaré inequality 77, 88
- polyhedral 171
- polyhedron 171, 350
- polynomial 396
- positive definite 209
- power method 436
- primal 358
- primal problem 320
- principle of virtual work 71, 141
- problem relaxed 362
- projected gradient 337
- propagation at finite velocity 257, 263
- Péclet number 6
  
- quadrature 157
- qualified constraint 312, 315
  
- Rayleigh quotient 216
- rectangular finite elements 191
- regular mesh 185
- regular open set 69
- regularizing effect 261
- regularity 129
- Rellich theorem 94
- reversible 8, 10, 60, 261
- rigid body motion 139
- Robin boundary condition 4
  
- saddle point 317
- Schrödinger’s equation 12
- second derivative 302
- self-adjoint 209
- separable 401
- simplex algorithm 353
- singularity 133
- slack variable 349
- Sobolev space 84
- space step, time step 14
- sparse matrix 421
- spectral decomposition 211
- spectral problem 214
- splitting 50
- stable 18, 37, 267
- state 372
- steady 11
- stencil 34

- stiff 268
- stiffness matrix 149, 151, 156, 225
- Stokes 13, 144
- strict convexity 290
- strong convexity 290
- strong solution 66
- subdifferential 392
- subgradients 392
- subordinate norm 406
- supergradient algorithm 393
- surface measure 68
- system 12
  
- test function 71
- trace theorem 90
- transmission 124
- transmission boundary
  - conditions 124
- transport 279, 368
- tree 381
- triangular finite
  - elements 171
- triangulation 171
- truncation error 35
  
- uniform mesh 154
- unimodular 365
- unisolvant 173, 192
- unstable 18
- upwind 16
- Uzawa's algorithm 338
  
- valuation 379
- variational formulation 71
- variational solution 113
- vertices 154, 381
- Von Neumann 41, 45, 272
  
- wave equation 9
- weak convergence 293
- weak derivative 81
- weak formulation 113
- weak solution 113
- well-posed 26

# Index of notations

$A^*$ or $A^t$ adjoint or transposed matrix .....	410
co $X$ convex envelope of $X$ .....	366
$C^k(\Omega), C^k(\overline{\Omega})$ space of $k$ times continuously differentiable functions .....	66
$C_c^\infty(\Omega)$ space of infinitely differentiable functions with compact support ....	82, 108
$C_c^\infty(\Omega)$ space of infinitely differentiable functions with bounded support .....	89
$\mathcal{D}(\Omega)$ space of infinitely differentiable functions with compact support .....	82
$\mathcal{D}'(\Omega)$ space of distributions .....	108
$\delta_{ij}$ Kronecker symbol .....	158
$dx$ (volume) Lebesgue measure .....	68
$ds$ (surface) Lebesgue measure .....	68
$\partial^\alpha$ partial derivative of order $\alpha$ .....	98
$\partial\Omega$ boundary of an open set $\Omega$ .....	68
$Epi(J)$ epigraph of a function $J$ .....	294
extr $K$ set of extremal points of a convex set $K$ .....	366
$H^1(\Omega)$ Sobolev space .....	86
$H_0^1(\Omega)$ trace free Sobolev space .....	90
$H^m(\Omega)$ Sobolev space of order $m$ .....	99
$H^{1/2}(\partial\Omega)$ trace space .....	94
$H(div)$ space of vector fields .....	103
$J'$ differential of a function $J$ .....	303
$L^2(\Omega)$ Lebesgue space .....	82
$L^\infty(\Omega)$ Lebesgue space .....	83
$n$ unit exterior normal .....	68
$ \mathcal{N} $ cardinal of a finite set $\mathcal{N}$ .....	378
$P_e$ entire envelop of $P$ .....	368
$\mathbb{P}_k$ space of polynomials of total degree $k$ or less .....	177
$\mathbb{Q}_k$ space of polynomials of degree $k$ or less in each variable .....	195
$\mathbb{R}_+^N$ half-space .....	92
$V'$ dual of a space $V$ .....	406